EARLY DETECTION OF ANTIMICROBIAL DRUG RESISTANCE FOR SALMONELLA TYPHI STRAINS USING MACHINE-LEARNING TECHNIQUES IN RWANDA

By

Benta Chepkirui

Reg. No: 219013791

A dissertation submitted in partial fulfillment of the requirements for the degree of Master of Science in Data Science (Biostatistics) in College of Business and Economics

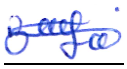
African Centre of Excellence in Data Science, University of Rwanda

Supervisor: Dr. Emil Ivan Mwikarago

September 2020

DECLARATION

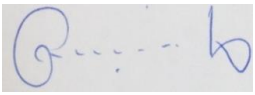
I declare that this dissertation is the result of my own work and has not been submitted for any other degree at the University of Rwanda or any other institution.

Benta Chepkirui 

Candidate Name Signature

Date: 29th September 2020

This MSc Thesis has been submitted for examination with my approval as dissertation advisor:

Dr. Emil Ivan MWIKARAGO 

Supervisor Signature

Date of submission 29th September 2020

ACKNOWLEDGMENT

I thank the Almighty God for enabling me to complete this thesis successfully. Indeed, it has been the most exciting experience at the African Centre of Excellence in Data Science. Besides, I wish to acknowledge the kind people who wholeheartedly contributed their invaluable assistance throughout this study.

First, my gratitude goes to my supervisor, Dr. Emil Ivan MWIKARAGO for believing in me all through and always providing me with endless support in my study. He always patiently provided me with unfailing support, inspiration and immense knowledge, especially in the medical field. I cannot thank him enough since without his extensive assistance, I would not have completed this study.

Most importantly, I wish to convey my sincere appreciation to the Centre Director, Dr Charles Ruranga and the Head of Teaching, Dr.Kabano for their continued guidance, moral support and motivation.

My regards also go to Prof. Cyprien Baribwira, Dr. Athanase Kiromera, Dr. Jackson Sebeza and Mr. Charles Mutai, my mentors for always supporting me through motivation and guidance whenever I needed.

Special thanks are to my entire family whose support, patience and love propelled and enabled me to achieve this success easily. I wish also to appreciate my classmates and friends, who shared with me their precious time and support. God bless you all.

ABSTRACT

Introduction: Antimicrobial drug resistance to Salmonella Typhi is among complex risk factors for morbidity and mortality, thus of global public health concern. Antimicrobial resistance patterns for S.Typhi suggest that currently, it presents a growing problem for developing countries. Early treatment of infectious diseases like Salmonella Typhi is key to combating high morbidity and mortality rates. Antimicrobial drug susceptibility test takes more than 24 hours, which is time-consuming and inefficient. More recently, predictive models have been used elsewhere to predict antimicrobial drug resistance patterns using machine-learning techniques for quick turn-around time and more efficient especially for patients in acute care conditions.

Objective: The main aim of this study was to predict a patient's antimicrobial drug resistance to Salmonella Typhi using Machine Learning Techniques.

Methodology: A cross-sectional study (2015 -2019) was conducted and of the 152 Salmonella Typhi isolates included in the study, 140 (92.1%) were from blood while 12 (7.9 %) were from stool. The Kirby-Bauer testing method was used for antimicrobial susceptibility. This study also predicted a patient's antimicrobial drug resistance to Salmonella Typhi using four machine-learning techniques namely; Support Vector Machine, Decision tree, Random Forest, and Logistic Regression using Antimicrobial Resistance data from a national reference laboratory on 765 cases in Rwanda. 5-fold cross-validation, classification report and confusion matrix metrics were used for performance measurement of the models.

Results: From 2015 to 2019, Cotrimoxazole resistance (86.2%) was highest compared to other first-line drugs: Ampicillin (85.5%) and chloramphenicol (80.9%). Nalidixic acid resistance (59.9%) and ciprofloxacin (20.4%) were high. There was lower resistance ceftazidime (32.9%), Tetracycline (9.9 %), and Cefotaxime (7.2%). All the built models had high predictions of antimicrobial drug resistance to Salmonella Typhi. Decision tree gave f1-score [0.89], accuracy [0.85] and AUC [0.82], Random forest gave f1-score [0.86], accuracy [0.90] and AUC [0.83], logistic regression gave f1-score [0.86], accuracy [0.88] and AUC [0.87] while Support Vector Machine f1-score [0.86], accuracy [0.89] and AUC [0.88]. However, a comparison that is based on the detailed performance measures suggests that the Support Vector Machine performs best.

Conclusion: There are significant antimicrobial resistance patterns in S.Typhi isolates to commonly used antibiotics. Applying machine-learning techniques can predict antimicrobial drug

resistance for Salmonella Typhi with high accuracy without clinical information. This approach may be extrapolated to predict antimicrobial drug resistance for any other organism. Further studies are recommended to determine the actual cost of predictive models on drug resistance in other clinical settings.

Keywords: AMR, Drug resistance, Salmonella Typhi, Machine Learning.

LIST OF ACRONYMS

CLSI - Clinical & Laboratory Standards Institute

XDR- Extensively Drug-resistant

AMR- Antimicrobial Resistance

MIC- Minimum Inhibitory Concentration

S.Typhi- Salmonella Typhi

LMIC- Low- and middle-income countries

WHO- World Health Organization

MDR- Multidrug resistance

WASH- Water-, sanitation-, and hygiene-related

FBD- Foodborne Disease Burden

NAL-R- Nalidixic Acid Resistant

ML- Machine Learning

TPR- True Positive Rate

FPR- False Positive Rate

AST- Antibiotic Susceptibility Testing

ICU- Intensive Care Unit

RF: Random Forest

ROC: Receiver Operating Characteristic

SVM: Support Vector Machine

LR: Logistic Regression

DT: Decision Tree

TABLE OF CONTENTS

TABLE OF FIGURES

LIST OF TABLES

CHAPTER ONE: INTRODUCTION

1.1 Background

Salmonella Typhi (S.Typhi) is a type of bacterium that causes a deadly infection called Typhoid fever. Its transmission can be via contaminated food or water [1] and is known to be among the major causes of death from Foodborne Diseases (FBD) [2]. World Health Organization (2019) predicts a global estimate of 11 to 21 million cases and over 128,000 to 161 000 typhoid-related death annually. Interestingly, Typhoid fever imposes more burden to Low- and middle-income countries (LMICs) than in high-income countries. It is estimated that 17.8 million cases of typhoid fever occur each year in LMICs [3] compared to 2 million cases that become infected annually in the united states of America and about 23 000 that die of such infections in similar settings [4].

Infectious diseases such as Typhoid have constantly been a threat to humanity, but antibiotics have saved millions of lives of people with these diseases. Since Alexander Fleming discovered antibiotics in 1928, they have significantly contributed to the reduction in mortality and morbidity rates associated with infectious diseases. However, in the past two decades, Antimicrobial resistance, (AMR) has spread outpacing the rate at which new antibiotics are developed. Research shows that more than 2 million people become infected on an annual basis with resistant bacteria to antibiotics in the United States while more than 23 000 die of similar infections [4]. Besides the high mortality rate, it is straining many countries financially due to increasing hospital admissions and drug usage[5].

AMR to S.Typhi is a complex risk factor for morbidity and mortality thus a worldwide public health concern. In LMICs, 72% and 10% are said to be with the burden of AMR as a result of S.Typhi for southeast Asia and Africa respectively [6]. Like any other curable disease, early treatment of infectious diseases like Salmonella Typhi is the key to combating the high mortality rate. However, antibiotic antimicrobial drug tic resistance has made treatment complicated. According to [7] AMR in S.Typhi may persist and eventually could lead to treatment failure.

Various studies have been conducted across Africa to determine the patterns and trends of antibiotic resistance to S.Typhi and changes are observed on an annual basis[8]. [9] Conducted a study in Nigeria from 1996 to 2008 and found high increasing resistance to the first-line drugs in Lagos. From 1996 to 2008, Ampicillin resistance to S. Typhi rose from 81.8% to 100% and that

remained constant to 2015, chloramphenicol increased from 63.6% to 100% and cotrimoxazole from 54.6% to 100%. Similarly, tetracycline increased from 63.6 to 100%. In Kenya, [10] conducted a study on 144 S.Typhi isolates. The findings revealed high resistance to the first-line drugs cotrimoxazole (70%), ampicillin (72%), and chloramphenicol (72%). 6% of the isolates were completely non-susceptible while 69% had intermediate susceptibility. Moreover, susceptibility to cefotaxime was at (83%) while ceftriaxone, gentamicin, and amoxicillin-clavulanic acid were (94%) and (97%) and (81%) respectively. A similar study conducted in Zimbabwe found that resistance to both ampicillin and chloramphenicol were between 83.3 to 100% in all the years. Ciprofloxacin resistance increased from 2012 (0%) to 2017 (22%) while that of tetracycline increased from 2012 (11.0%) to 2017 (46.3%) [11].In Rwanda, a study found an increase in Multidrug-resistant S.typhi (MDRST) in 2007 to 2018 from 18.2% to 52.8% for Ampicillin, 18.2% to 25% for chloramphenicol and 18.2% to 50% for cotrimoxazole [12]. Moreover, Salmonella Typhi isolates were not resistant to Ciprofloxacin, Ceftriaxone, and Levofloxacin. The decreased susceptibility to Nalidixic acid from 97% to 80.5% suggested that Fluoroquinolones are tending to be no option. The resistance patterns keep changing annually and therefore this study aims at revealing the 2015-2017 patterns in Rwanda.

Alongside obtaining patterns through simple descriptive statistics as other studies in Rwanda, this study stretches out to apply Machine-learning techniques on the available data to determine whether a patient with S.Typhi is resistant or not to an antibiotic. The most common method used for Antibiotic Susceptibility Testing (AST) is the Kirby Bauer Method [13].

Application of machine learning in predicting resistance is one of the ways that can contribute effectively to the proper use of antibiotics. This is because one can determine whether a patient is resistant to an antibiotic on time before administering it, which, will be most useful in an ICU environment. Unlike the common method used for antibiotic susceptibility testing (Kirby Bauer Susceptibility Testing),[13] , that normally take 24 hours, [14], to determine the presence or absence of acquired resistance to a variety of antibiotics, application of a machine learning-based methodology enables earlier detection of resistance through prediction [15]. Moreover, the Kirby Bauer Susceptibility Testing method is not effective in prescribing treatment when a patient is in

a critical condition [16] but machine learning ensures the rational use of antibiotics and therefore can reduce mortality in ICU patients[17]

1.2 Problem Statement

In Africa, Multidrug Resistance (MDR) typhoid fever cases (resistance to all the existing first-line drugs: chloramphenicol, ampicillin, trimethoprim-sulfamethoxazole/ Cotrimoxazole), are still common [6],[18]. Besides, S.Typhi has developed resistance to even existing second-generation fluoroquinolones such as Ciprofloxacin, as a burden in Africa. Rwanda is no exception[12]. Third-generation cephalosporins are normally preferred drugs when a S.Typhi strain is resistant to fluoroquinolones [19]. [20] Argued that AMR in third-generation cephalosporins such as ceftriaxone, cefixime have proved to be rare in Africa. However, it is recently reported in Kenya that susceptibility to ceftriaxone was (94%) and cefotaxime (83%) [10]. This implies an existing resistance to third-generation cephalosporins. The spread of extensively drug-resistant (XDR) (resistance to chloramphenicol, ampicillin, trimethoprim-sulfamethoxazole, fluoroquinolones, and third-generation cephalosporin) is a real threat that may limit therapeutic options especially in critically ill patients [21]. This study will therefore contribute to the body of knowledge on AMR in Rwanda, and for third-generation cephalosporin S.Typhi resistance in particular.

Studies in Rwanda on antibiotic resistance to S.Typhi have only revealed the patterns and trends [12],[22] in antibiotic resistance using simple descriptive statistics but none have used the available data to predict antibiotic resistance in S.Typhi using Machine Learning Techniques. The idea behind this is to determine whether a patient with S.Typhi is resistant or not to an antibiotic on real-time. Clinicians in the ICU have long relied on Antibiotic susceptibility tests such as the Kirby Bauer Method and minimum inhibitory concentration (MIC) methods to determine the susceptibility of antibiotics yet acquiring results takes 24 hours or more after the sample is collected[23]. Rwanda also uses the methods that take up to 48 hours for incubation to get susceptibility results[24]. Also, the use of genome-sequencing data to determine susceptibility remains limited and may take time [25]. This time can however be shortened by employing machine learning techniques [26]. Using machine learning techniques to predict antibiotic susceptibility in patients with critical illnesses can help to achieve high performance [27]. Machine

Learning Techniques provide a way to use different models such as random forest to predict antibiotic resistance where such metrics as accuracy and precision are used to choose the best performing model [28]. Notably, using readily available data can sometimes help in prescribing treatment to patients in critical condition [16]. In this era of high technology utilization for different purposes, there are reports in the literature where artificial intelligence has been used to innovatively solve public health issues such as drug resistance [29][30]. Studies that explore the mechanism of using artificial intelligence (AI) through machine learning tools with multiple models for predicting drug resistance have not been fully documented in data science platforms, thus creating a big knowledge gap.

Therefore, in addition to revealing the patterns in AMR to S.Typhi using simple descriptive statistics like in existing studies, this study introduces the use of AI on readily available data in a reference laboratory by building machine-learning models that will best predict antimicrobial drug resistance and specifically for S.Typhi burden in Rwanda.

1.3 Objectives

This study's primary objective was to find out if applying machine-learning techniques on available S.Typhi AMR data (2015-2019) in a reference laboratory could predict antimicrobial resistance without relevant clinical information.

The secondary objectives are as follows:

1. To determine the antimicrobial resistance pattern for Salmonella Typhi among commonly used antibiotics at a reference laboratory in Rwanda
2. To develop antimicrobial resistance predictive models for drug susceptibility patterns using Machine Learning Techniques

1.4 Significance and justification of the study

This research is significant because it will reveal the current patterns of AMR in S.Typhi for Rwanda Reference Laboratory. Through this, it will propel the urgency for typhoid preventive measures and the formulation of rational interventions for the reduction of the burden of resistance. Using machine-learning techniques may predict early detection of antimicrobial drug resistance to reduce treatment failure and mortality in intensive care unit (ICU) patients in real-time. It will also reduce the length and cost of the hospitalization of the patient. Besides, the machine-learning

model for predicting AMR will be an important tool for clinicians to anticipate the resistance of bacteria infection, thus give an antibiotic type suitable for prescription of the appropriate patient. It will limit antibiotic misuse to reduce the prevalence of antimicrobial-resistant bacteria and therefore relieve Rwanda, the financial strain caused by the increased hospital admissions and drug usage. Finally, it will create a roadmap for the concerned health institutions to come up with timely decisions in terms of preventive public health policies and vaccination priorities in LMICs.

1.5 Scope and limitation of the Study

This research's focus was on building machine-learning models for predicting antimicrobial resistance of S.Typhi strains using the available AMR data (20115-2019) from a Reference Laboratory in Rwanda. The Machine Learning Classifiers built were decision tree and random forest, Support Vector Machine and logistic regression

CHAPTER TWO: LITERATURE REVIEW

2.1 Patterns of AMR to Salmonella Typhi

Presently, there is a challenge in establishing the pattern of AMR to establish policies and prevention protocols. The WHO is encouraging governments to publish data regularly to enable clear intervention with a paucity of evidence [31]. Nonetheless, scarce evidence on AMR exists due to limited global AMR data where above 40 percent of Africa has limited AMR data [8], including Rwanda [32]. However, with the limited data, informative AMR patterns can be obtained to implement policies that combat the burden of resistance

2.1.1 Resistance to Fluoroquinolones

Although according to [33], the first-line drugs seem to work well, (cotrimoxazole resistance (6.1%) and chloramphenicol resistance (13.8%), while Ceftriaxone and azithromycin resistance was 16.1 and 5.78% respectively), various studies have found otherwise. Fluoroquinolones such as Ciprofloxacin are the most reliable treatment of typhoid fever because of the rise of AMR to standard first-line drugs [34]. It has been found that Ciprofloxacin resistance in Salmonella Typhi is rare and regarding them as the drugs of choice [8].

On the contrary, other studies reveal that S.Typhi has recently proved to be resistant to ciprofloxacin and even show a decreased susceptibility [12,13,33]. An example is a study in the United States, where among the NAL-R Typhi isolates tested, 99% were showed a decreased susceptibility to ciprofloxacin [38]. Similar findings were found by [39],[40] whereof 169 isolates from travelers to Pakistan, 133 (79%) were fluoroquinolone non-susceptible. More studies reveal high resistance to ciprofloxacin[41][42]. [42], conducted a research between 2012-2014 on 1979 (69%) S. Typhi and 893 (31%) S. Paratyphi. S.Typhi resistance to Ciprofloxacin decreased from 94% in 2012 to 88% in 2014. Despite the decrease, it still maintained the high resistance.

The same applies to Asia, where research shows a constant increase in Nalidixic acid and fluoroquinolones resistance from 20% in 2001–2005 to 65% in 2011–2015. [38], observed changing patterns in enteric fever. 750 (69%) out of 1872 Typhi isolates had either reduced susceptibility to ciprofloxacin or were resistant to nalidixic acid. 99% of the S.Typhi isolates that were resistant to nalidixic acid showed an increased resistance or completely non-susceptible to ciprofloxacin. Intermediate susceptibility to ceftriaxone was observed from a traveler to India.

There was no resistance to azithromycin. In Rwanda, fluoroquinolones have proved not to be reliable [12]. Even though the same study showed that Ciprofloxacin was completely sensitive to S.Typhi, there was reduced resistance to nalidixic acid.

2.1.2 Resistance to third-generation cephalosporins

Third-generation cephalosporins have proved to be an option when fluoroquinolones are resistant [15]. According to a research conducted by [19], on patterns of antimicrobial sensitivity of Salmonella typhi on 16 isolates, 10 (62.5%) were MDR, 03(18.75%) were resistant to Ciprofloxacin and Azithromycin and none was resistant to Ceftriaxone and Ceftazidime. Interestingly, Nalidixic acid resistance was at 100%. This suggests that third-generation cephalosporins are an option when fluoroquinolones are resistant to S.Typhi.

Additional studies have found low resistance in third-generation cephalosporins and high resistance to first-line drugs and fluoroquinolones. For example [43], found that out of 431 S.Typhi isolates, 28.3% isolates were MDR while resistance to ampicillin was 96/335 (28.7%), chloramphenicol 115/430 (26.7%), and cotrimoxazole 117/431 (27.1%). Also, resistance to nalidixic acid was 92.3% and had intermediate sensitivity to ciprofloxacin. No isolate was resistant to cefixime and ceftriaxone while only seven isolates were resistant to azithromycin. Similarly, [44] examined 200 isolates of salmonella infections where most were 142 (71%) S.Typhi followed by Salmonella Paratyphi A 58(26%). There was no resistance to cefepime, four isolates were resistant to ceftriaxone while Only four of the isolates were MDR. However, 48% were resistant to ciprofloxacin. [10] Conducted a study on 144 S.Typhi isolates where resistance to the first-line drugs was high at cotrimoxazole (70%), ampicillin (72%) and, chloramphenicol (72%). 6% of the isolates were completely non-susceptible while 69% had intermediate susceptibility. Moreover, susceptibility to cefotaxime was at (83%) while ceftriaxone, gentamicin, and amoxicillin-clavulanic acid were (94%) and (97%) and (81%) respectively. Unfortunately, it has been found that S.Typhi resistance to third-generation cephalosporins is increasing. An example is a study in Asia that found resistance to third-generation cephalosporin rose from 1.5% in the period from 2006 through 2010 to 4% in the period from 2011 through 2015 [7]. This is alarming since it leads to XDR [40]. According to [45], 76% of the 239 S.Typhi were MDR while resistance to ciprofloxacin was 91%. Interestingly, 48% of the isolates were XDR and thus Meropenem and azithromycin were the options.

Examining a study, [46], conducted on 223 (81.1%) S. Typhi isolates and 52 (18.9%) S. Paratyphi isolates, it reveals that S.Typhi had high resistance under the third-generation cephalosporins. Focusing only on the S.Typhi isolates, six (2.6%) isolates were MDR typhoid while two (0.9%) isolates were XDR typhoid. Cefixime was 60.9% susceptible while cefotaxime was 65.8% susceptible. Resistance to ciprofloxacin was at 49.9%. A different interesting study found that the risk of infections associated with ceftriaxone-resistant *S* Typhi is bigger among children aged 15 years and younger[1]. This means that AMR to S.Typhi differs across different age groups and it may be of interest to consider determining patterns of S.Typhi resistance across different age-groups.

2.1.1 Multidrug resistance (MDR) and extensively drug-resistance (XDR) to S.Typhi

Studies suggest that MDR cases in the United States are associated with traveling to India [39],[40][47]. Research by [47] found that out of the two hundred seventy-two (13%) isolates that were MDR, 85% had traveled to the Indian subcontinent. LMIC reveals existing and even increasing multidrug resistance (MDR) to Salmonella Typhi [48],[12],[49]. To remedy this, a study decided to screen herbal plants [50]. Research in Asia, [33], found low MDR in North-India through 2011–2017 at 2.73%. Despite the low MDR resistance this Asia, it is of great interest to be aware of the emergence of XDR Typhi [39],[40]. For example, [46] conducted a study on 223 (81.1%) S. Typhi isolates and 52 (18.9%) S. Paratyphi isolates to find out the patterns in Antimicrobial Susceptibility. Focusing on the S.Typhi isolates, six (2.6%) isolates were MDR typhoid while two (0.9%) isolates were XDR typhoid.

A recent systematic review conducted by [7] found that in Africa, there existed high MDR resistance of above 90% while in Asia, there was a small number of MDR strains of S.Typhi equal to less than 20% between 2011 and 2015. Moreover, the MDR proportion was decreasing over the same period. There exist MDR in Kenya[10] and even Rwanda[12]. This growing problem could lead to XDR thus limits treatment options and may be of interest to find out the current patterns in Rwanda.

2.1.2 Annual Patterns

Resistance to S.Typhi in Africa tends to change an annual basis and mostly increasing[8]. For example, in Nigeria [9] , conducted a study from 1996 to 2008 and found high increasing resistance to the first-line drugs in Lagos. From 1996 to 2008, Ampicillin resistance to S. Typhi rose from 81.8% to 100% that remained constant to 2015, chloramphenicol increased from 63.6% to 100% and cotrimoxazole from 54.6% to 100%. Similarly, tetracycline increased from 63.6 to 100%. A similar study conducted in Zimbabwe found that resistance to both ampicillin and chloramphenicol were between 83.3 to 100% in all the years. Ciprofloxacin resistance increased from 2012 (0%) to 2017 (22%) while that of tetracycline increased from 2012 (11.0%) to 2017 (46.3%) [11].

In Rwanda, a study found an increase in Multidrug-resistant S.typhi (MDRST) from 3/33(9.1%) in 2017 to 9/36(25%) in 2018. From 2017 to 2018, first-line drugs increased from 18.2% to 52.8% for Ampicillin, 18.2% to 25% for chloramphenicol and 18.2% to 50% for cotrimoxazole [12]. Moreover, Salmonella typhi isolates were not resistant to Ciprofloxacin, Ceftriaxone, and Levofloxacin. The decreased susceptibility to nalidixic acid from 97% to 80.5% suggested that Fluoroquinolones are tending to be no option. The resistance patterns keep changing annually and therefore this study aims at revealing the 2015-2017 patterns in Rwanda.

2.2 AMR Stewardship

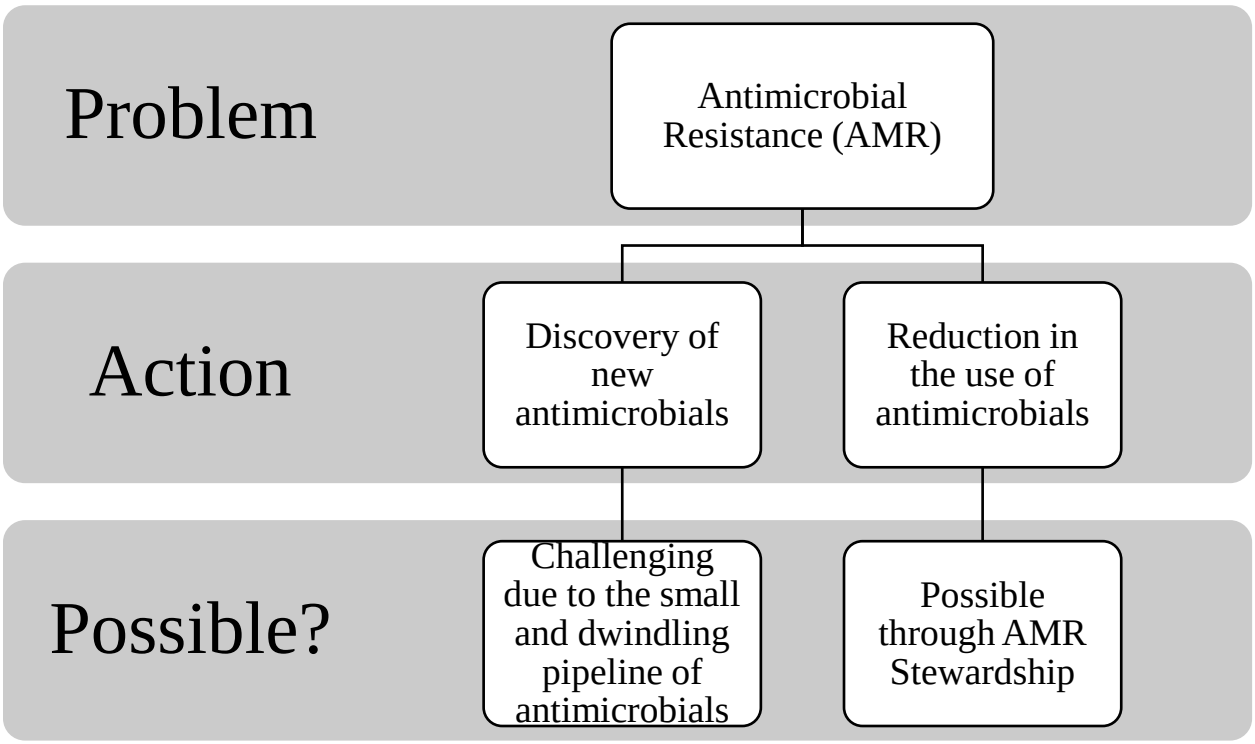


Figure 2.2.1 AMR Stewardship[51]

According to WHO [51], Antimicrobial Stewardship is the systematic effort to educate and persuade prescribers of antimicrobials to follow evidence-based prescribing to stem antibiotic overuse and thus antibiotic resistance. AMR results when microorganisms like bacteria, fungi, parasites, and bacteria change when they are exposed to antimicrobials. This means that the continued use of antimicrobials is the main driver of AMR.

Figure 2.2.1 illustrates that to effectively mitigate this threat, either new drugs should be discovered or the use of antimicrobials can be reduced. However, it is unlikely that the solution will come from the discovery of new drugs due to the already dwindling pipelines of antimicrobials and the slow discovery of drugs that are not toxic to humans. Therefore, Antimicrobial Stewardship is the only practical and current solution by finding ways of using the available antimicrobials correctly.

2.3 Use of Machine Learning Techniques in Predicting Antimicrobial Resistance

Data Scientists have strived to work with clinicians to improve health using machine learning including prediction of antibiotic resistance [52]. The use of Machine Learning Techniques using

whole-genome sequencing for antibiotic susceptibility testing (WGS-AST) has proved to be more powerful than culture-based susceptibility testing [53].

Machine learning with SVM and the LR algorithms achieved high sensitivity (95-100%) by using genomic data to predict antibacterial susceptibility in Mycobacterium tuberculosis [54]. More metrics like AUCs have been used and high performance is achieved. For example, [55] conducted a similar study using repeated cross-validation and found that the average AUC (0.979) for first-line drugs and AUC (0. 936) for second-line drugs for the highest performing models. Average accuracy has also been used as a metric and it achieved high performance in prediction of antibiotic resistance in Escherichia coli and A. pleuropneumonia ,respectively, from large-scale pan-genome data different models [56] [57]. Nontyphoidal *Salmonella* genomes have also been used to predict MICs for antibiotics using XGboost with high accuracy of 95% [58]

Apart from genomic data, it is argued that the application of patients' clinical history and demographics also performs well in predicting antibiotic resistance [55]. Moreover [59] considered demographics, clinical and patient history in building machine learning models (Gradient Boosting Decision Trees (GBDT) and Logistic Regression) which gave great predictive power.

However,[16], assures that available data in a laboratory such as patients' demographic factors, data from cultures, and susceptibility testing without any clinical data on patient's history can provide reliable predictions on antibiotic susceptibility that helps clinicians in choosing appropriate antibiotic therapy. Therefore, this study finds out if readily available data on AMR in a reference laboratory can be used to predict antibiotic resistance in S.Typhi using Machine Learning Techniques.

The use of AI technology through machine learning techniques in the prediction of antibiotic resistance has become popular recently [30]. From the literature, authors have used machine learning to develop prediction models in various domains including health care, modeling among other areas [60]. These techniques have also been used to predict antibiotic resistance of various bacteria using demographic and clinical factors [28]. According to [27] and [61], various factors including age are important in predicting AMR. It is therefore important to explore the use of machine learning techniques to find out if readily available data on AMR in Rwanda reference laboratory could be used to predict antibiotic resistance for S.Typhi.

CHAPTER THREE: METHODOLOGY

This section elaborates novel approaches that aim to eliminate some limitations of the susceptibility testing tools that are currently used in laboratories. With this, therefore, not only patterns on S.Typhi resistance to commonly used antibiotics are revealed but also the application of AI through machine-learning techniques to predict resistance are introduced.

3.1 Research design and Source of data

A five-year Cross-sectional study was conducted from 2015 to 2019. The isolates of this study were obtained from Rwanda reference laboratory located at Kigali, Rwanda. The laboratory is responsible for supporting health service at all levels of health care by developing policies regulating laboratories in Rwanda, training laboratory personnel, supervising laboratories, and providing external quality control of health facilities in Rwanda. The target population was isolated from patients of all age groups who had a positive culture of S.Typhi bacteria (2015-2019).

3.1.1 Sampling process

The isolates obtained were from blood and stool. Originally, raw data had 188 *S.typhi* isolates. After data cleaning, which included the dropping of records with missing data, duplicates, and negative blood cultures, *S.typhi* isolates 152 obtained. Each isolate was tested to at most 8 antibiotics.

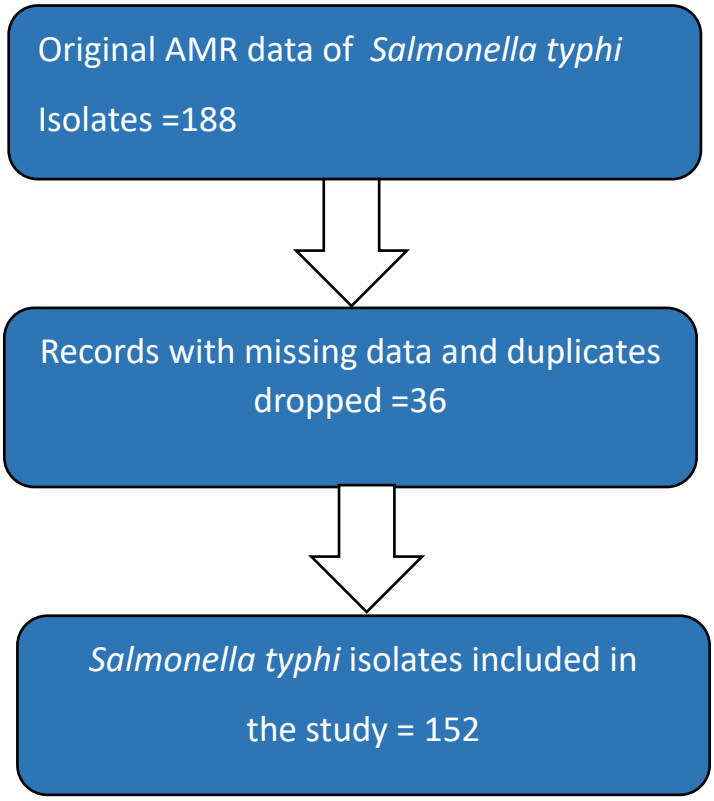


Figure 3.1.1 Sampling Process

3.1.2 Antibiotic sensitivity testing

Blood cultures are normally sent to the microbiology laboratory in culture medium bottles. The incubation period was 7 days at 35 degrees Celsius. Antibiotic sensitivity testing was done by disk diffusion (Kirby-Bauer testing). An interpretation was done per the Clinical and Laboratory Standards Institute (CLSI) standards. The following standard steps are followed to find out if a patient's isolate is resistant to a bacterium using the Kirby Bauer Method:

1. Any type of patient sample is obtained, say blood, urine, or swab depending on the location of the suspected infection.

2. The sample is then cultured in the laboratory by placing it in an ideal environment in a laboratory to ensure cell growth of the bacteria in the sample). Markedly, most species of bacteria are the same that it is not easy to distinguish them by use of the microscope only.
3. Obtaining an antibiogram, (overall profile of antimicrobial susceptibility testing results of a specific microorganism to a set of antimicrobials) for the positive cultures. This usually takes 24 hours or more. The resistance and susceptibility are then set according to Minimum inhibitory concentration (MIC), which is the lowest concentration of an antimicrobial that will inhibit the visible growth of an organism in an ideal growing condition. It is the gold standard for determining susceptibility [62].

3.1.3 Data Description

The data in this study included the patients' demographic characteristics: age group (categorical), gender (categorical), health facility (categorical) and province location (categorical), ID, the name of the organism, antibiotics (categorical), type of the specimen (categorical) and the results (binary). The outcome variable was the results that showed whether an isolate was resistant or sensitive to a certain antibiotic.

Table 3.1.1 Data Description

No.	**Variable Name**	**Description**
1	AGE	Patient's age in years
2	GENDER	The type of sex (female or male) of the patient from whom the isolate was extracted
3	PROVINCE	The province from which the patient lives
4	HEALTH_FACILITIES	The Health facility that the patient visited

5	ANTIBIOTICS	The commonly used antibiotics for S.Typhi that were used to test for susceptibility
6	SAMPLE_TYPE	The type of sample that was obtained from a patient (blood and stool)
7	LOGIN_YEAR	The year that the patient visited the health facility
8	RESULTS	The results from the susceptibility test of the S.Typhi isolates (resistant or sensitive) to an antibiotic

3.1.4 Dealing with Categorical Data

Label encoding of the target variable was done (Resistant=1 or Sensitive=0). Dummies were set to convert the independent categorical variables into a form that enabled the ML algorithms to do a better job in prediction.

3.2 Descriptive Analysis

Descriptive analysis was performed to understand the data in terms of frequencies and patterns. It involved determining the following:

- Frequency of S.Typhi isolates from each age group
- Age-wise distribution of S. Typhi.
- The contribution of each sample type (blood and stool samples)
- Annual percentage resistance of S. Typhi towards various antimicrobials
- Overall Resistance to each antibiotic in the five years

3.3 Building Machine Learning model

3.3.1 Feature Selection

A filter method of feature selection (Spearman's correlation) was used to determine the relevant features for building a machine-learning model to predict antibiotic resistance in S.Typhi. This was done by the use of a heat map. In addition to multicollinearity, it showed the correlation between results (dependent variable) and independent variables. From these, all the independent variables were relevant since there was no strong multicollinearity. Moreover, from literature, these variables are predictors of AMR [63].

3.3.2 Machine learning Models Used

The python was used to build the machine learning classifiers for predicting antibiotic resistance in S.Typhi. Decision tree, random forest, Support Vector Machine, and logistic regression were used to predict patient-specific antibiotic resistance to S.Typhi.

3.3.3 Decision Tree

Decision Tree [47] is a machine learning model used for both classification and regression problems. A decision tree is a graphical representation of all the possible answers to come up with a decision as illustrated in figure 2.3.1. This model is created through induction [64], where the trees are built and Pruning where the unnecessary structure is removed from a decision tree to reduce overfitting and ease interpretation.

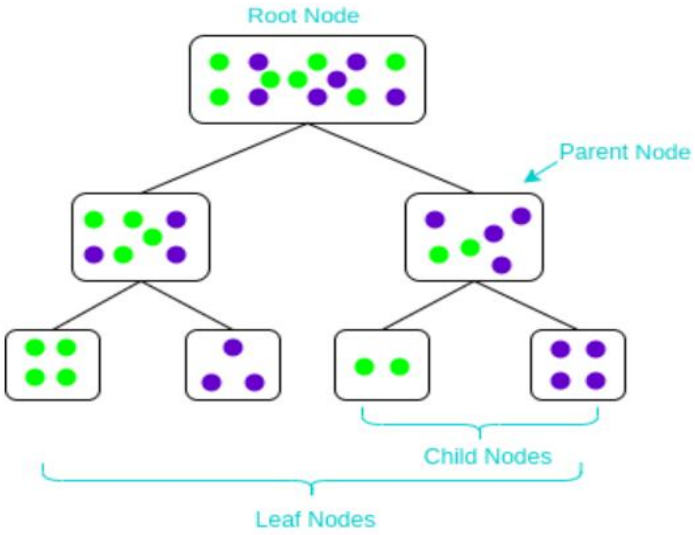


Figure 3.3.1 Basic Decision Tree Terminologies [65]

3.3.4 Random Forest Classifier

Random forest is a tree-based algorithm that entails the construction of numerous trees (decision trees), then joining their output to better the generalization ability of the model. The method of joining these trees is referred to as an ensemble method. It is the process of combining weak learners (specific trees) to produce a strong learner [66]. Random Forest Classifier performs well and corrects for decision trees routine of overfitting to their training set [67].

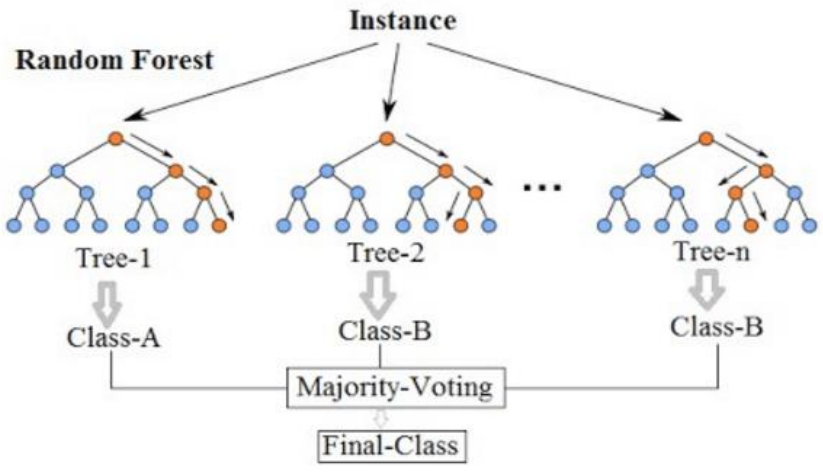


Figure 3.3.2 Random Forest Classifier [68]

3.3.5 Logistic Regression

Logistic Regression is a machine-learning algorithm that is used during classification problems. It is part of the generalized linear models' family and suitable when the output is binary (0 or 1). By default, it returns the set of probabilities of the target class. Logistic regression makes the assumptions that the dependent variable need to binomial distributed, there is a linear relationship between the explanatory variables and the link function (logit) and the response variable must have mutually exclusive and exhaustive categories.

How Logistic Regression works

Just as Linear Regression's equation is:

$$Y = b_0 + b_1 x + Error \qquad \textit{Equation 1}$$

The Logistic function is obtained by:

$$P(Y = 1/X) = \frac{e^{(b_0 + b_1 x)}}{e^{(b_0 + b_1 x)} + 1} \qquad \textit{Equation 2}$$

$$(x + a)^n = \sum_{k=0}^{n} \binom{n}{k} x^k a^{n-k} \qquad \textit{Equation 3}$$

$$P(x) = \frac{e^{(b_0 + b_1 x)}}{e^{(b_0 + b_1 x)} + 1} \qquad \textit{Equation 4}$$

$$P(e^{(b_0 + b_1 x)} + 1) = e^{(b_0 + b_1 x)} \qquad \textit{Equation 5}$$

$$P = e^{b_0 + b_1 x} - P.e^{(b_0 + b_1 x)} \qquad \textit{Equation 6}$$

$$P = e^{(b_0 + b_1 x)}(1 - P) \qquad \textit{Equation 7}$$

$$\frac{P}{1-P} = e^{(b_0 + b_1 x)} \qquad \textit{Equation 8}$$

$$In(\frac{P}{1-P}) = b_0 + b_1 x$$ *Equation 9*

This is the **logit** function that assumes a sigmoid function with a range of probabilities between 0 and 1 as shown in figure 2.3.3.

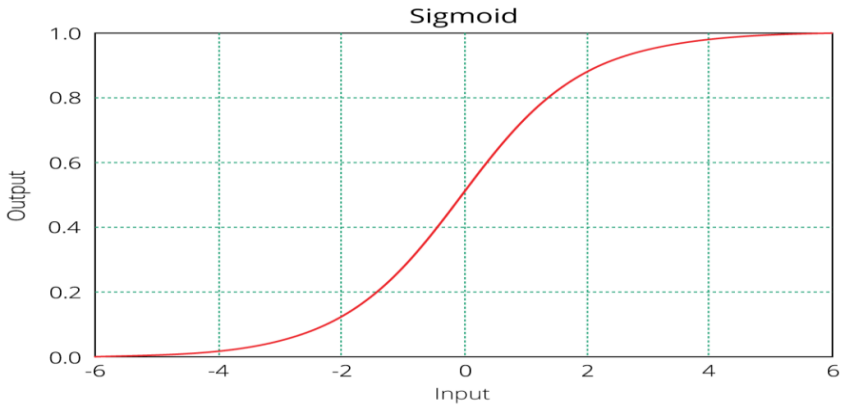


Figure 3.3.3 Logistic Regression [69]

3.3.6 Support Vector Machine (SVM)

A support vector Machine (SVM) is a binary classification machine-learning algorithm classifies a dataset in the best possible way. The distance between either nearest points is the margin and distinct the two classes of data points, several possible hyperplanes could be selected. These decision boundaries assist in classifying the data points. Thus, the objective is to choose a hyperplane with the maximum possible margin between support vectors in the given dataset. The purpose of maximizing the margin distance is to give some reinforcement so that later, data points can be classified with more confidence.

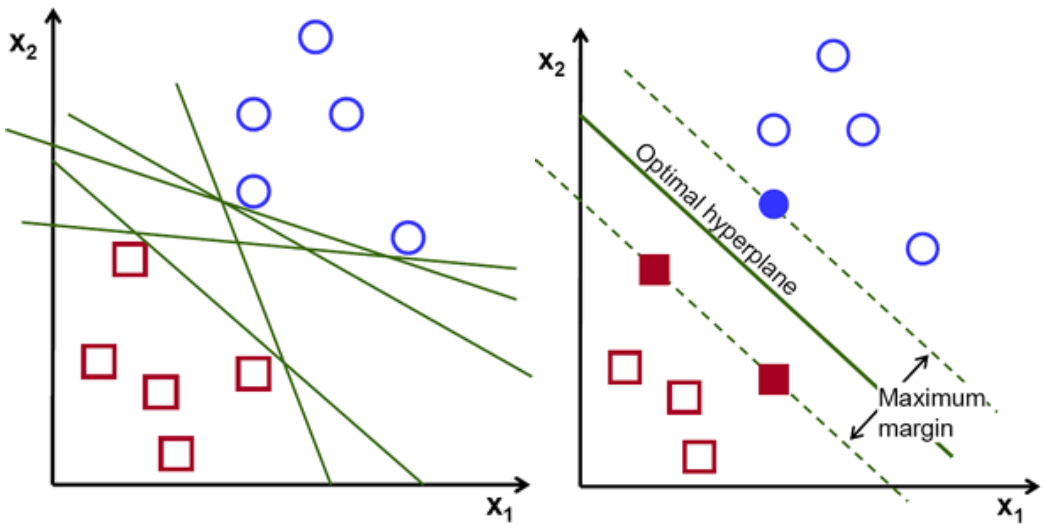


Figure 3.3.4 Support Vector Machine reference[70]

3.4 Model Evaluation

Each model was evaluated using a classification report (the recall, precision, F1-score, and accuracy) from repeated 5-fold cross-validation, confusion matrix, and Area under ROC Curve.

3.4.1 5-fold cross-validation

K-fold cross-validation is a resampling procedure that is used to evaluate machine-learning models normally when the data is limited. Therefore, this procedure was best for our study. The idea is to use a limited sample to estimate how the model will be expected to perform in general on data that was not used when the training of the model was performed. It also produces less biased models because it makes sure that all observations from the data are used during training and testing.

K-fold cross-validation works by splitting the dataset into k-subsets without replacement. For instance, in this study, a 5-fold cross-validation evaluation method was used where the data were randomly split into five mutually exclusive subsets with nearly equal sizes. In the first evaluation iteration, the first fold was used to test the model and the other was used to train the model. In the second evaluation iteration, the second fold was used to test while the rest as the training set. The testing operation was then repeated 5 times. It implies that at this point, each fold of the 5 folds had been used as the testing set and 5 models were obtained. For each instance in the dataset, accuracy score was determined and an overall, accuracy estimate was provided.

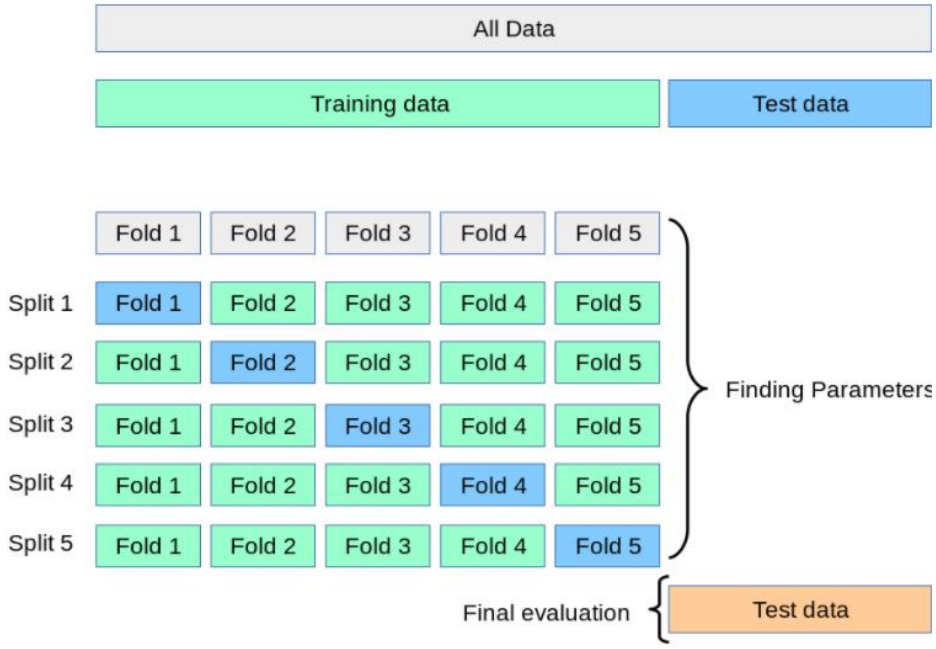


Figure 3.4.1 5-fold cross- validation reference[71]

3.4.1 Confusion Matrix

A confusion matrix is used to evaluate the quality of the output of a classifier. The diagonal elements are the number of points where the predicted label is equal to the true label. The off-diagonal on the other hand, are the labels that the classifier mislabeled. It is better when the diagonal values of the confusion matrix are higher since it indicates that many predictions are correct.

Table 3.4.1 Confusion Matrix

		Predicted Class	
		Class= Yes	Class=No
Actual Class	Class=Yes	True Positive	False Negative
	Class=No	False Positive	True Negative

True Positives tells us the number of cases that the classifier correctly predicted that the person is positive. In our case, it tells us the number of cases that the classifier correctly predicted that the person is resistant

True Negative tells us the number of cases that the classifier correctly predicted that the person is negative, in this study it tells us the number of cases that the classifier correctly predicted that the person is sensitive.

False Positives are also called Type 1 error. It tells us the number of cases that the classifier incorrectly predicted that the person is positive when in fact they are negative. For this study, it tells us the number of cases that the classifier incorrectly predicted that the person is resistant when in fact is sensitive

False Negatives, on the other hand, is also called Type II error and tells us the number of cases that the classifier incorrectly predicted that the person is negative when in fact are positive. Thus, in our study, it is the number of cases that the classifier incorrectly predicted that the person is sensitive when in fact is resistant

From the confusion matrix, we can obtain evaluation metrics that will enable us to compare different machine learning models according to their performance. These metrics include:

- Sensitivity|Recall|True Positive Rate
- Specificity
- False Positive Rate
- Precision

3.4.2 Precision and Recall

3.4.2.1 Precision

This is the ability of a classifier not to label as positive a sample that is negative. In other words, it is the percentage of instances labeled as positive that is positive. It leads us to know, out of all the samples that the classifier predicted as positive, what portion of it was correct. Precision is obtained by:

$$Precision = \frac{TP}{TP+FP} \qquad \textit{Equation 10}$$

3.4.2.2 Recall

Recall shows the ability of the classifier to find all the positive values. It leads us to know, out of all the positive samples, what portion of did my classifier pick up. It is given by:

$$Recall = \frac{TP}{TP+FN} \qquad \textit{Equation 11}$$

3.4.3 F1-Score

F1 Score is a blend of precision and recall

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \qquad \textit{Equation 12}$$

3.4.4 Classification Accuracy

Classification accuracy is a measurement that tells us how best a machine learning model can identify patterns between variables based on training data. If a model can generalize to 'unseen' data better, it implies that it can produce better predictions.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \qquad \textit{Equation 13}$$

3.4.5 The Area under ROC Curve

A ROC curve is a plot that shows the performance of a binary machine learning classifier, as its discrimination threshold is changed. It is obtained by plotting the True Positive Rate (**TPR**), which is also called Sensitivity versus False Positive Rate (**FPR**), which can also be obtained by (1-Specificity). Area Under ROC Curve (AUC) makes it easier to compare one ROC Curve to another.

Sensitivity answers the question when the actual value is positive, how often is the prediction correct? It is also known as recall or True Positive Rate (TPR) and is obtained by:

$$TPR = \frac{TP}{TP+FN} \qquad \textit{Equation 14}$$

False Positive Rate answers the question: when the actual value is negative, how often is the prediction incorrect? Which is also 1-specificity and is obtained by:

$$FPR = \frac{FP}{FP+TN} \qquad \textit{Equation 15}$$

3.5 Hyperparameter Tuning

The performance of most machine learning methods highly depends on the model architecture defined by the hyperparameter settings. Several methods exist for automatically optimizing hyperparameters, such as grid search, random search, informed search, Bayesian optimization, etc. All of these methods of hyperparameter tuning require a "search space" that specifies the range of possible parameters to evaluate during optimization. While some of the automated methods can take a long time to optimize, manual tuning is far worse since it involves a very tedious process that often results in the evaluation of unpromising "search space". Essentially, the primary objective is to search for hyperparameter settings that optimize an objective function, which yields the lowest cross-validation error. In our implementation, we used both a grid search to find the optimum set of hyperparameters that minimized cross-validation error.

Hyperparameter tuning is an external configuration to a model whose values cannot be estimated from data. A machine-learning model's performance depends on it. To do this, various methods exist that automatically optimize the hyperparameters. For instance, for the Support Vector Machine, the parameters are called support vectors whose examples are the kernel, C, and gamma.

The most commonly used are grid search, random search, and informed search. They all require a "search space" that specifies the range of possible parameters to evaluate during optimization. Some of these automated methods take a long time to optimize while manual tuning takes more time since it involves a very tedious process that mostly leads to unreliable "search space".

The main objective is to attain the optimal hyperparameters whose objective function yields the lowest cross-validation error. In this study, a grid search was used to find the optimum set of hyperparameters that minimized cross-validation error.

CHAPTER FOUR: RESULTS

4.1 Epidemiological characteristics and distribution of isolates

Table 4.1.1 shows that of the 152 isolates included in the study, 140 (92.1%) were from the blood while 12 (7.9 %) were from the stool. The annual distribution of the isolates was 28 (18.4 %) in 2015, 22 (14.5 %) in 2016, 50 (32.9 %) in 2017, 21 (13.8%) in 2018 and 31 (20.4 %) in 2019. The isolates were from individuals with a mean age of 16.9 with a standard deviation of 13.1. Half 76 (50.0%) of the isolates were children between 0-15 years, 55 (36.2 %) between 16-30 years 18 (11.8 %) between 31-45 years and 3 (2 %) above 45 years. Most of the isolates were from male 86 (56.6%) while the 66 (43.4%) were from female

Geographically, most isolates were from Burera district 46 (30.3%) then Butaro 45 (25.6%), Kibungo DH 19 (12.5%). However, some districts such as Rwamagana, Kibuye DH, Nyarugenge, Rubavu and, Huye did not have any isolates 0 (0.0%). Looking at the health facilities, most isolates were from Butaro 45(29.6%), followed by Kibungo DH 19(12.5%), then Ngarama 14(9.2%), and Byumba 12(7.9%). Other health facilities had a very low number of isolates such as Shyita DH, Nyamasheke, Nyagatare, Cyanika (NYAMAGABE) that only had 1 (0.7%) isolates. Kibuye DH had none 0 (0.0%)

Table 4.1.1 Epidemiological characteristics and distribution of isolates

		TOTAL (n=152)
Factor		**Value**
Age	Mean	16.9 (13.1)
	Median[Min, Max]	15.5 [0.00,77]
Age Category	0-15 Years	76 (50.0%)
	16-30 Years	55(36.2%)
	31-45 Years	18(11.8%)
	46 and above years	3(2.0%)
Gender	Male	86(56.6%)
	Female	66(43.4%)
Sample Type	Blood_Culture	140 (92.1%)
	Culture_Stool	12(7.9%)
Year	2015	28(18.4%)
	2016	22(14.5%)
	2017	50(32.9%)
	2018	21(13.8%)
	2019	31(20.4%)
District	Bugesera	3(2.0%)
	Burera	46(30.3%)
	Gakenke	0(0%)
	Gasabo	2(1.3%)
	Gatsibo	14(9.2%)
	Gicumbi	12(7.9%)
	Gisagara	0(0%)
	Huye	0(0%)
	Kamonyi	0(0%)
	Karongi	7(4.6%)
	Kayonza	0(0%)
	Kicukiro	0(0%)
	Kirehe	16(10.5%)
	Muhanga	0(0%)
	Musanze	2(1.3%)
	Ngoma	21(13.8%)
	Ngororero	0(0%)
	Nyabihu	1(0.7%)
	Nyagatare	1(0.7%)
	Nyamagabe	1(0.7%)
	Nyamasheke	4(2.6%)
	Nyanza	2(1.3%)
	Nyarugenge	0(0%)
	Rubavu	0(0%)
	Ruhango	3(2.0%)
	Rulindo	4(2.6%)
	Rusizi	4(2.6%)
	Rutsiro	9(5.9%)
	Rwamagana	0(0%)

4.2 Antibiotic Resistance to S.Typhi

4.2.1 Patterns of Antibiotic Resistance to S.Typhi

Table 4.2.1 Trends of Antibiotic Resistance to S.Typhi

	2015 (n=28)	2016 (n=22)	2017 (n=50)	2018 (n=21)	2019 (n=31)	Total (n=152)
Ampicillin						
Resistant	24 (85.7%)	16 (72.7%)	46 (92.0%)	19 (90.5%)	25 (80.6%)	130 (85.5%)
Sensistive	4 (14.3%)	6 (27.3%)	4 (8.0%)	2 (9.5%)	6 (19.4%)	22 (14.5%)
Cefotaxime						
Resistant	0 (0%)	1 (4.5%)	9 (18.0%)	0 (0%)	1 (3.2%)	11 (7.2%)
Sensistive	28 (100%)	21 (95.5%)	41 (82.0%)	21 (100%)	30 (96.8%)	141 (92.8%)
Ceftazidime						
Resistant	1 (3.6%)	1 (4.5%)	26 (52.0%)	7 (33.3%)	15 (48.4%)	50 (32.9%)
Sensistive	27 (96.4%)	21 (95.5%)	24 (48.0%)	14 (66.7%)	16 (51.6%)	102 (67.1%)
Chloramphenicol						
Resistant	20 (71.4%)	14 (63.6%)	46 (92.0%)	17 (81.0%)	26 (83.9%)	123 (80.9%)
Sensistive	8 (28.6%)	8 (36.4%)	4 (8.0%)	4 (19.0%)	5 (16.1%)	29 (19.1%)
Ciprofloxacin						
Resistant	1 (3.6%)	1 (4.5%)	11 (22.0%)	6 (28.6%)	12 (38.7%)	31 (20.4%)
Sensistive	27 (96.4%)	21 (95.5%)	39 (78.0%)	15 (71.4%)	19 (61.3%)	121 (79.6%)
Cotrimoxazole						
Resistant	21 (75.0%)	18 (81.8%)	43 (86.0%)	20 (95.2%)	29 (93.5%)	131 (86.2%)
Sensistive	7 (25.0%)	4 (18.2%)	7 (14.0%)	1 (4.8%)	2 (6.5%)	21 (13.8%)
Nalidixic						
Resistant	13 (46.4%)	4 (18.2%)	49 (98.0%)	16 (76.2%)	9 (29.0%)	91 (59.9%)
Sensistive	15 (53.6%)	18 (81.8%)	1 (2.0%)	5 (23.8%)	22 (71.0%)	61 (40.1%)
Tetracycline						
Resistant	1 (3.6%)	2 (9.1%)	6 (12.0%)	3 (14.3%)	3 (9.7%)	15 (9.9%)
Sensistive	27 (96.4%)	20 (90.9%)	44 (88.0%)	18 (85.7%)	28 (90.3%)	137 (90.1%)

From table 4.2.1, we see that the maximum number of S.Typhi isolates (n=50) were found in the year 2017, the year 2019 had 31 isolates, 2015 had 28 isolates while 2016 had 22 S.Typhi isolates and finally 2018 had the least number of 21 isolates of S.Typhi.

We observe that over 5 years (from 2015 to 2019), Cotrimoxazole was the most resistant among the first-line drugs 86.2% followed by Ampicillin 85.5% then chloramphenicol 80.9%. Resistance

to fluoroquinolones was relatively high since resistance to Nalidixic acid was at 59.9% while ciprofloxacin resistance was 20.4%. Resistance ceftazidime was 32.9% while resistance to Tetracycline was at 9.9 %. Cefotaxime appears to have the lowest resistance of 7.2%.

Focusing on annual resistance, 24 (85.7%) isolates were resistant to Ampicillin in 2015. The resistance decreased to 16 (72.7%) in 2016 and increased to 46 (92%) in 2017. In 2018, the resistance percentage slightly reduced to 19 (90.5 %) and eventually decreased to 25 (80.6 %) in 2019.

A similar pattern was observed in chloramphenicol where 20 (71.4%) of the isolates were resistant in 2015. This resistance decreased to 14 (63.6%) in 2016 and increased to 46(92%) in 2017. In 2018, S.Typhi resistance to chloramphenicol reduced to 17 (81.0 %) then it slightly increased to 26 (83.9 %). Unlike these, two first-line drugs, Cotrimoxazole resistance gradually increased from 21(75%) in 2015 to 18(81.8%) in 2016 and increased again to 43(86 %) in 2017 and 20 (95.2%) in 2018. There was a slight reduction to 29 (93.5 %) in 2019. Resistance to Ciprofloxacin also increased slightly from 1 (3.6 %) in 2015 to 1 (4.5 %) in 2016 and sharply increased to 11(22 %) in 2017. This resistance persistently increased to 6 (28.6 %) in 2018 and 12 (38.7 %) in 2019. Nalidixic acid showed a decrease in resistance from 46.4% in 2015 to 18.2% in 2016 then increased highly to 98% in 2017 then reduced to 16 (76.2 %) in 2018 and eventually reduced again to 9 (29.0%) in 2019. Resistance to Ceftadizime was 3.6% in 2015. This resistance slightly increased to 4.5% in 2016 and greatly increased to 52% in 2017. A decrease of 7 (33.3 %) was observed in 2018 and an increase of 15 (48.4 %) in 2019. There was no resistance to cefotaxime in 2015. However, in 2016, it had a low resistance of 1 (4.5%) which increased to 9 (18%) in 2017 and reduced to 0 (0%) in 2018 which increased to 1 (3.2 %) in 2019. Resistance to Tetracycline was low but gradually increased from 1 (3.6%) in 2015 to 2 (9.1%) in 2016, 6 (12%) in 2017 and 3 (14.3%) in 2018. This resistance reduced slightly to 3 (9.7 %) in 2019.

4.2.2 MDR and XDR Antibiotic Resistance

Table 4.2.2 MDR AND XDR

	2015 (n=28)	2016 (n=22)	2017 (n=50)	2018 (n=21)	2019 (n=31)	Total (n=152)
MDR						
1	19 (67.9%)	14 (63.6%)	41 (82.0%)	14 (66.7%)	20 (64.5%)	108 (71.1%)
Missing	9 (32.1%)	8 (36.4%)	9 (18.0%)	7 (33.3%)	11 (35.5%)	44 (28.9%)
XDR						
1	0 (0%)	0 (0%)	4 (8.0%)	0 (0%)	0 (0%)	4 (2.6%)
Missing	28 (100%)	22 (100%)	46 (92.0%)	21 (100%)	31 (100%)	148 (97.4%)

From table 4.2.2, we observe that over five years, 2015-2019, MDR to S.Typhi is at an average of 108/152 (71.1%). 19/ 28 (67.8%) S.Typhi isolates were MDR in 2015. This percentage decreased in 2016 where 14/22(63.6%) of the isolates were MDR. In 2017, the percentage increased drastically to 41/50 (82%). In 2018, MDR reduced to 14/21 (66.7%) which reduced to 20/31 (64.5 %)

XDR on the other hand is at an average of 4/152 (2.6%) through 2015-2019. It was not observed in both 2015 and 2016, 2018 and 2019 however, it is alarming that 4/50 (8%) of the S.Typhi strains were XDR in 2017.

4.2.3 Age-related Patterns of Antibiotic Resistance to S.Typhi

Table 4.2.3 Age-related Patterns of Antibiotic Resistance to S.Typhi

	0-10 Years (n=60)	11-20 Years (n=45)	21-30 Years (n=26)	Above 30 Years (n=21)	Total (n=152)
Ampicillin					
Resistant	52 (86.7%)	36 (80.0%)	23 (88.5%)	19 (90.5%)	130 (85.5%)
Sensistive	8 (13.3%)	9 (20.0%)	3 (11.5%)	2 (9.5%)	22 (14.5%)
Cefotaxime					
Resistant	4 (6.7%)	3 (6.7%)	2 (7.7%)	2 (9.5%)	11 (7.2%)
Sensistive	56 (93.3%)	42 (93.3%)	24 (92.3%)	19 (90.5%)	141 (92.8%)
Ceftazidime					
Resistant	22 (36.7%)	14 (31.1%)	6 (23.1%)	8 (38.1%)	50 (32.9%)
Sensistive	38 (63.3%)	31 (68.9%)	20 (76.9%)	13 (61.9%)	102 (67.1%)
Chloramphenicol					
Resistant	47 (78.3%)	38 (84.4%)	21 (80.8%)	17 (81.0%)	123 (80.9%)
Sensistive	13 (21.7%)	7 (15.6%)	5 (19.2%)	4 (19.0%)	29 (19.1%)
Ciprofloxacin					
Resistant	15 (25.0%)	8 (17.8%)	2 (7.7%)	6 (28.6%)	31 (20.4%)
Sensistive	45 (75.0%)	37 (82.2%)	24 (92.3%)	15 (71.4%)	121 (79.6%)
Cotrimoxazole					
Resistant	53 (88.3%)	37 (82.2%)	22 (84.6%)	19 (90.5%)	131 (86.2%)
Sensistive	7 (11.7%)	8 (17.8%)	4 (15.4%)	2 (9.5%)	21 (13.8%)
Nalidixic					
Resistant	35 (58.3%)	26 (57.8%)	17 (65.4%)	13 (61.9%)	91 (59.9%)
Sensistive	25 (41.7%)	19 (42.2%)	9 (34.6%)	8 (38.1%)	61 (40.1%)
Tetracycline					
Resistant	8 (13.3%)	1 (2.2%)	2 (7.7%)	4 (19.0%)	15 (9.9%)
Sensistive	52 (86.7%)	44 (97.8%)	24 (92.3%)	17 (81.0%)	137 (90.1%)

We can see from table 4.2.3 that the number of isolates reduces as the age groups go higher. Of those between 0-10 years (n=60) which reduced to n=45 among those from 11-20 years. Those of age group 21-30 years and above 30 years had a sample size of n=26 and n=21 respectively.

It is interesting from table 3.2.3 that S.Typhi's resistance to cotrimoxazole 53 (88.3 %) was the highest among the age group 0-10 years. Ampicillin, chloramphenicol, and Nalidixic acid also had a high resistance of 52 (86.7%), 47 (78.3%), and 35 (58.3%) respectively in the same age group. Ceftazidime had a resistance of 22 (36.7%) while Ciprofloxacin had a relatively lower resistance of 15 (25%). Tetracycline had a low resistance of 8 (13.3%) while the lowest resistance was observed to be from Cefotaxime 4 (6.7 %).

At age 11-20, resistance to all the antibiotics dropped except for Chloramphenicol which increased from 47 (78.3%) to 38 (84.4%). Despite the drop, resistance to Cotrimoxazole 37 (82.2%) and Ampicillin 36 (80.0 %) was still high in this age group. Resistance to both Nalidixic acid 26 (57.8 %) and, Ceftadizime 14 (31.1%) was relatively high. Ciprofloxacin, Cefotaxime and Tetracycline had relatively low resistance of and 8 (17.8 %), 3 (6.7%), and 1 (2.2%) respectively.

Among those of age 21-30, there was an increase in resistance of all antibiotics compared to those of age 11-20 except for chloramphenicol, Ciprofloxacin and Ceftadizime that dropped from 38 (84.4%), 8 (17.8 %), 14 (31.1%) to 21 (80.8 %), 2 (7.7 %), 6 (23.1 %) respectively. Ampicillin had the highest resistance of 23 (88.5%) followed by Cotrimoxazole 22 (84.6 %). Nalidixic resistance was also high in this age group 17 (65.4%). Unlike other age groups where Tetracycline, Cefotaxime, and Ciprofloxacin had the lowest resistance of 2 (7.7 %) each.

Of those above 30 years, cefotaxime had the lowest resistance of 2 (9.5%). Resistance to Tetracycline, Ciprofloxacin, and Ceftadizime increased to 4 (19%), 6 (28.6%), and 8 (38.1%) respectively. Despite the drop in Nalidixic resistance, the resistance was still high at 13 (61.9%). Just as other age groups, Ampicillin and Cotrimoxazole had a very high resistance of 19 (90.5%) while Chloramphenicol also had a high resistance of 17 (81.0%).

4.3 Distribution of the target variable

Fig 4.3.1 shows the distribution of the target variable (RESULTS) that determines whether one is resistant or sensitive to an antibiotic is bivariate (Resistant=1 or Sensitive=0). The sensitive cases were 455 (59.5%) while resistant cases, on the other hand, were 310 (40.5%). This means that our data set is balanced.

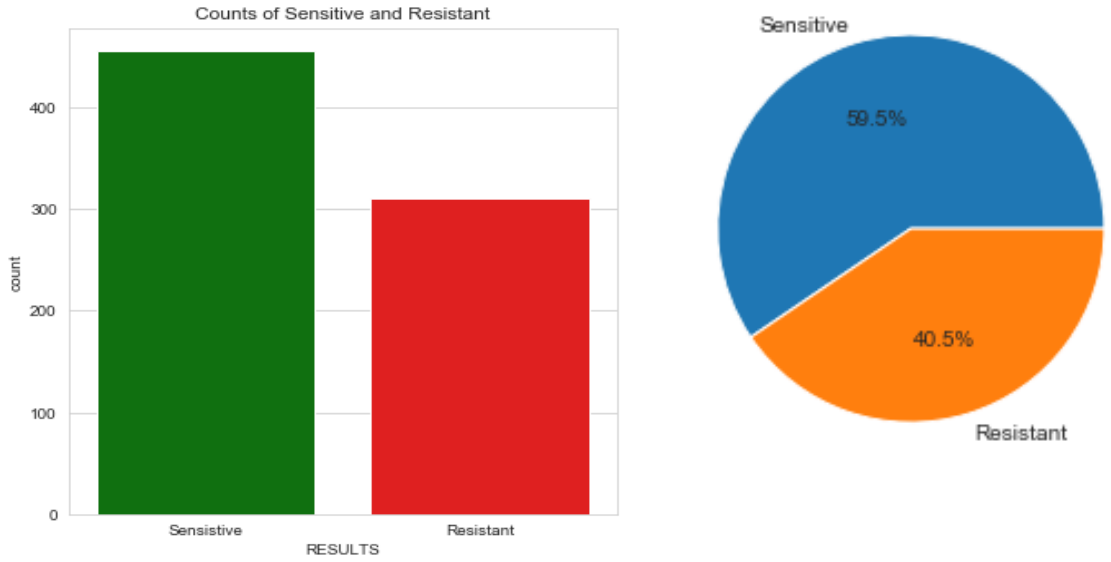


Figure 4.3.1 Distribution of the target variable

4.4 Feature Selection: Heat map

Correlation matrices help us to understand our data in detail in terms of the extent of correlation. From the output in figure 4.4.1, we can see that the correlations between each independent variable to the target variable (results) are clear. For instance, login_year tends to be negatively correlated to results by -0.22 while antibiotics is positively correlated to results by 0.12. Moreover, the level of multicollinearity can be detected. They are relatively small, like gender and antibiotics that correlate by -0.00089.

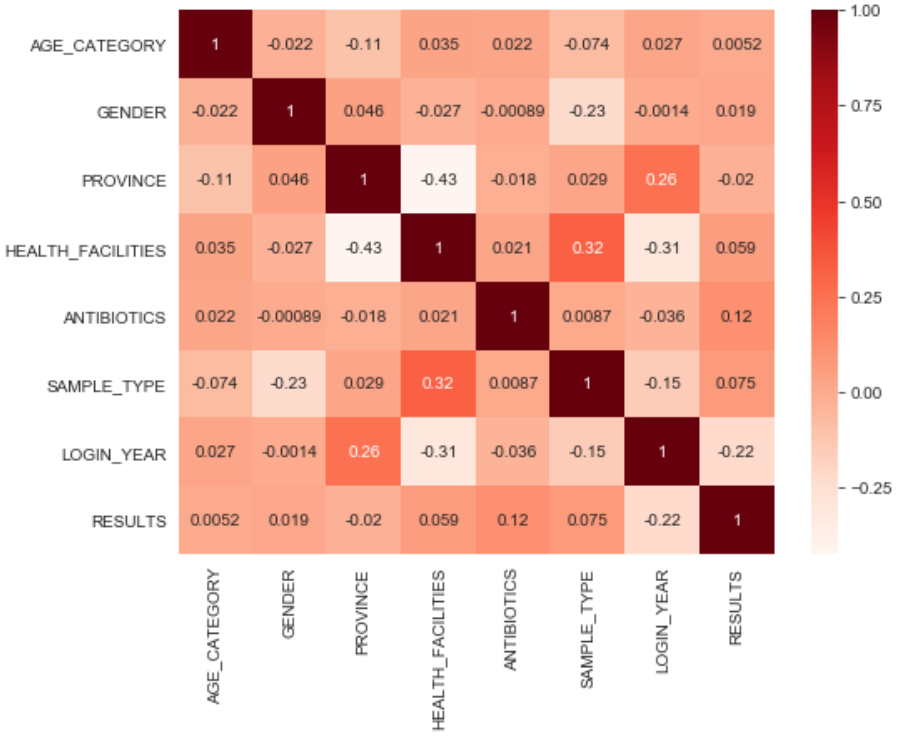


Figure 4.4.1 Feature Selection: Heat map

4.5 Evaluation Scores of the Machine Learning Models

4.5.1 Support Vector Machine

Table 4.5.1. Support Vector Machine

Model	Precision Score	Recall Score	F1_Score	Accuracy
Support Vector Machine	0.86	0.9	0.89	0.85

Results shown in *table 4.5.1* show that the Support Vector Machine had a recall of 0.9, which was higher than the precision (0.86). These two lead to an F1_Score of 0.89. It is also of importance to note that the accuracy of 0.85 was lower than the F1_Score. *Fig 4.5.1.1,* on the other hand, shows that the SVC Area Under the ROC Curve was 0.8781.

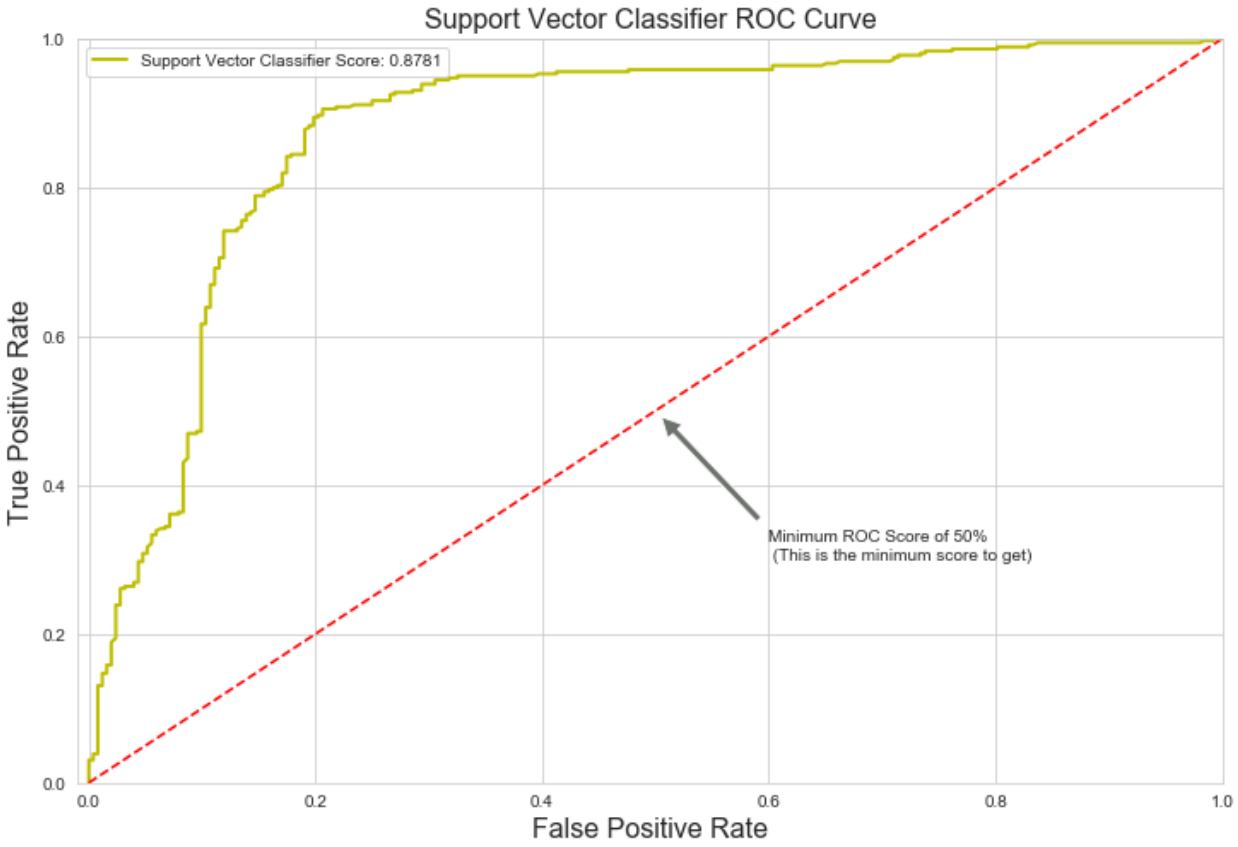


Figure 4.5 Support Vector Machine ROC curve

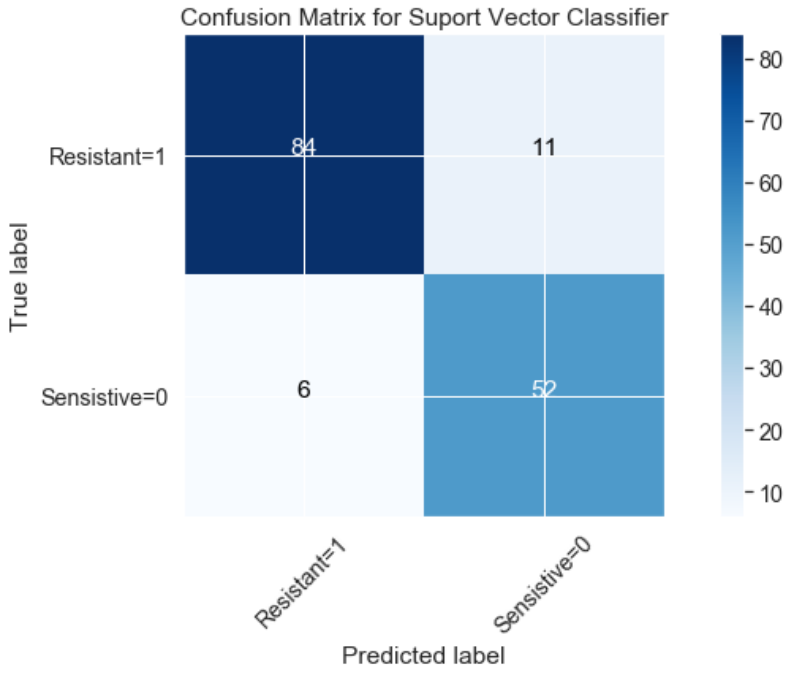


Figure 4.5 Support Vector Machine Confusion Matrix

From the confusion matrix of Support Vector Machine in figure 4.5.1.2, we observe that the True positives are 84, the True negatives are 52, false positives are 6 and the False Negatives are 11.

4.5.2 Random Forest

Table 4.5.2 Random Forest

Model	Precision Score	Recall Score	F1_Score	Accuracy
Random Forest	0.85	0.88	0.86	0.90

From table 4.5.2, Random Forest Classifier had an accuracy of 0.9, which was higher than its F1_Score (0.86). The Recall (0.88) was higher than the precision (0.85). The AUC of the Random Forest (fig 4.5.2.1), was 0.8280.

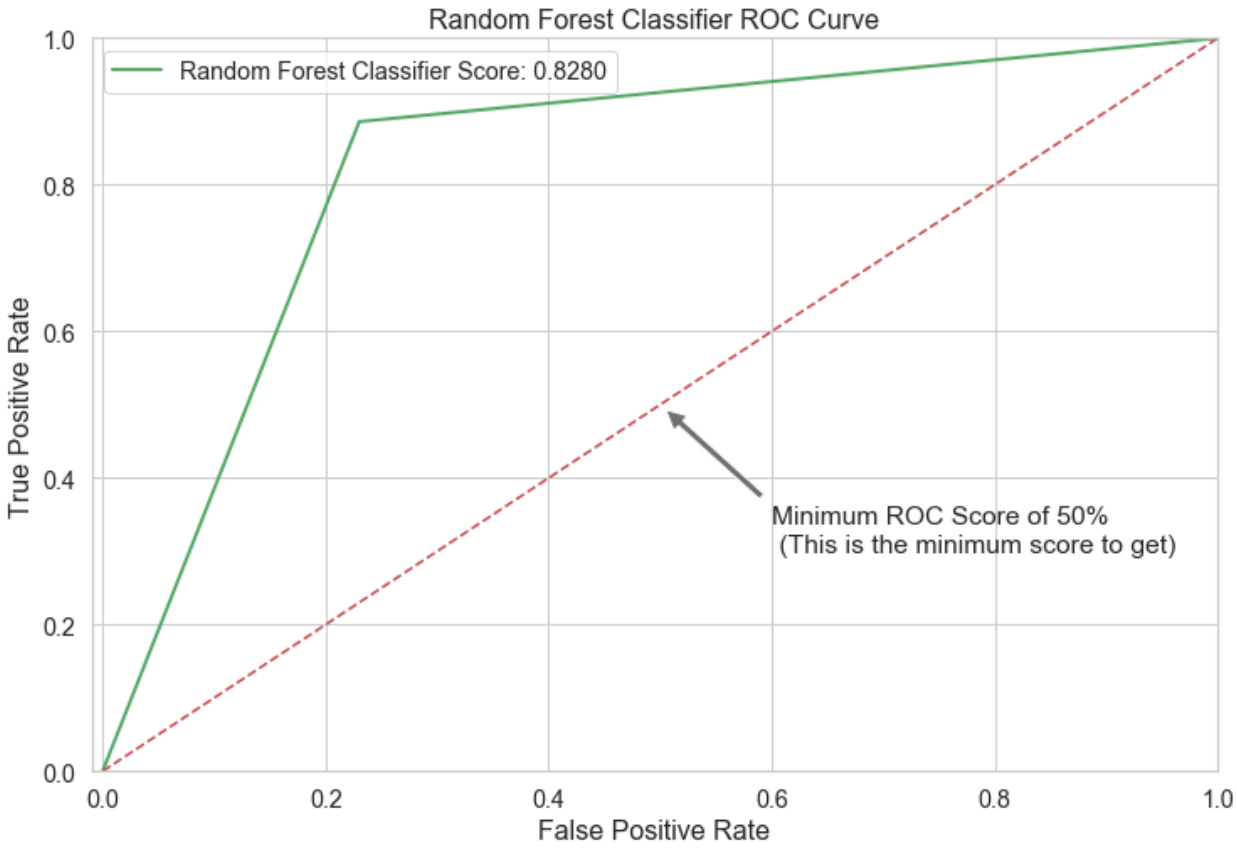


Figure 4.5 Random Forest Classifier ROC Curve

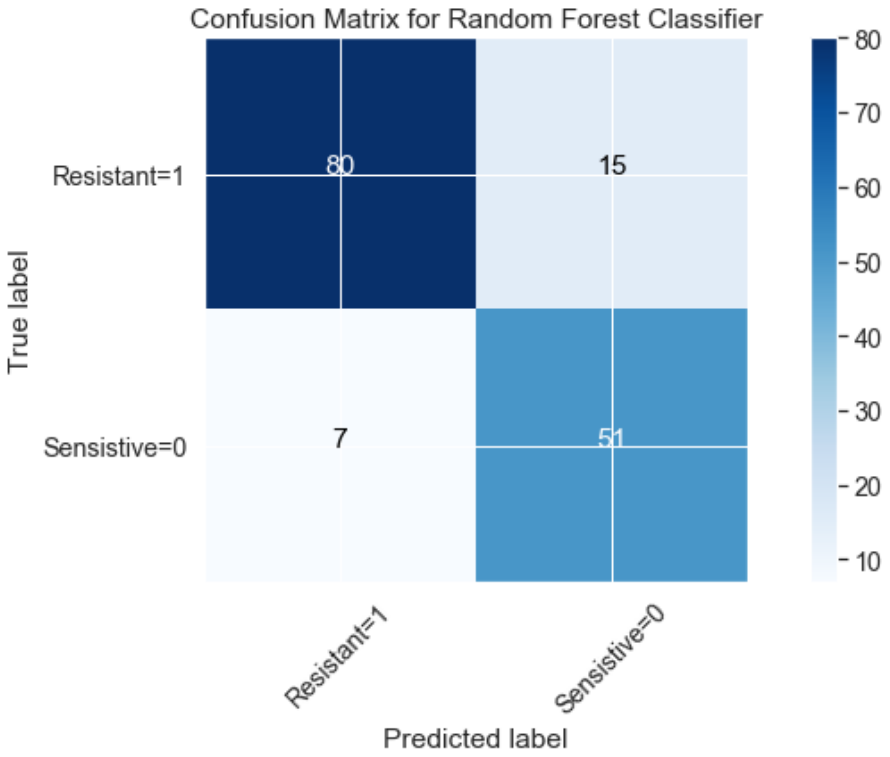


Figure 4.5 Random Forest Classifier confusion Matrix

From the confusion matrix Random Forest Classifier in figure 4.5.2.2, we observe that the True positives are 80, the True negatives are 51, false positives are 7 and the False Negatives are 15.

4.5.3 Logistic Regression

Table 4.5.3 Logistics Regression

Model	Precision Score	Recall Score	F1_Score	Accuracy
Logistic Regression	0.83	0.89	0.86	0.88

We observe in (*table 4.5.3*) that Logistic regression's recall score was the highest (0.89), followed by the accuracy score (0.88). The precision score was the lowest among (0.83) all the scores

leading to a low F1_Score (0.85) compared to the accuracy. *Fig 4.5.3.1* reveals that the Logistic Regression AUC was high at 0.8736

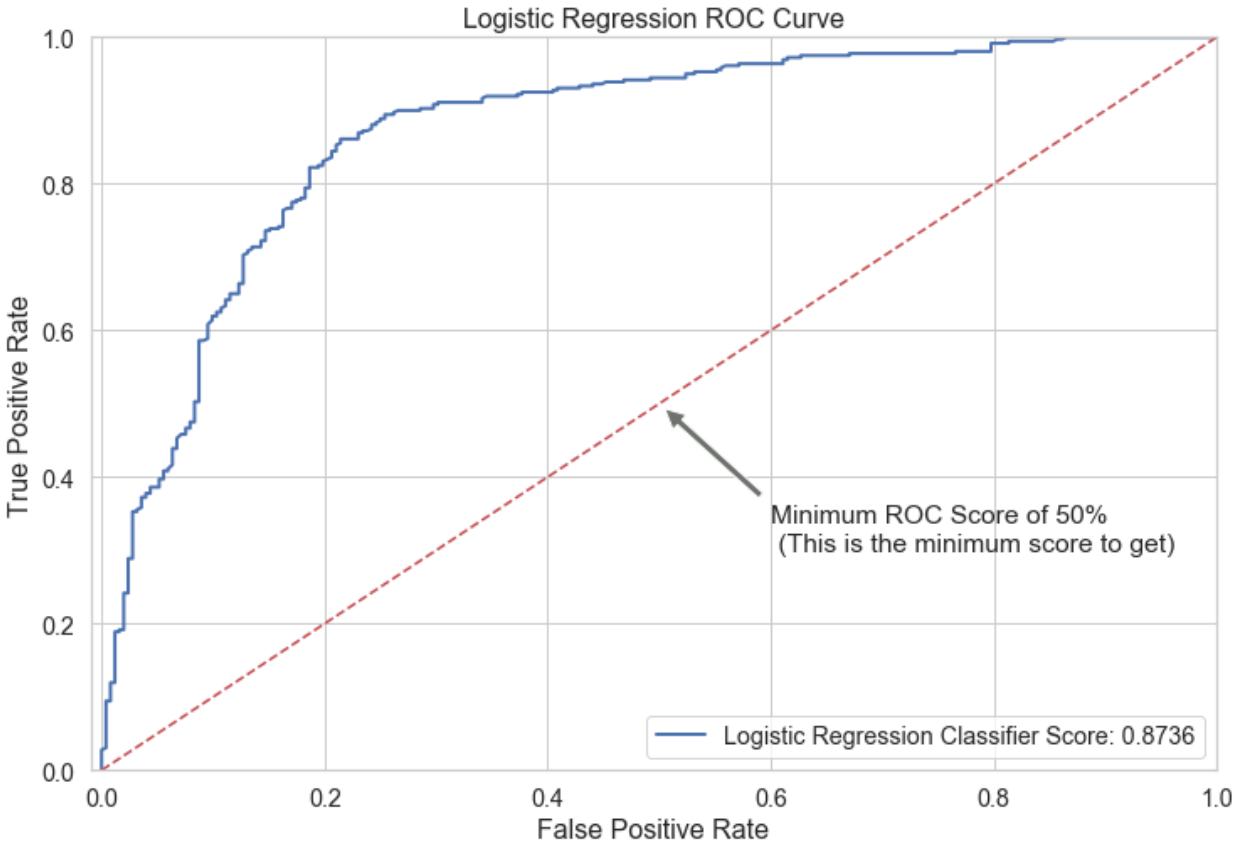


Figure 4.5 Logistic Regression ROC Curve

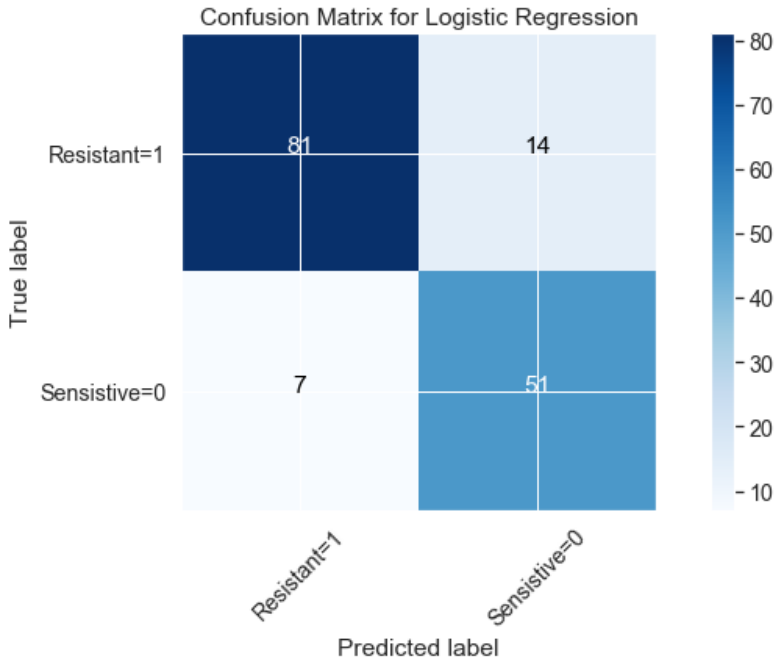


Figure 4.5 Logistic Regression Confusion Matrix

From the confusion matrix of the Logistic Regression Classifier in figure 4.5.3.2, we observe that the True positives are 81 the True negatives are 51, false positives are 7 and the False Negatives are 14.

4.5.4 Decision Tree

Table 4.5.4 Decision Tree

Model	Precision Score	Recall Score	F1_Score	Accuracy
Decision Tree	0.85	0.86	0.86	0.89

Decision Tree (table 4.5.4) had an accuracy score of 0.89, which was higher than all its metric scores. The precision score (0.85) was slightly lower than the recall score (0.86). This lead to an F1_Score of 0.86. The AUC on the other hand was low at 0.8228 as shown in *fig 4.5.4.1*

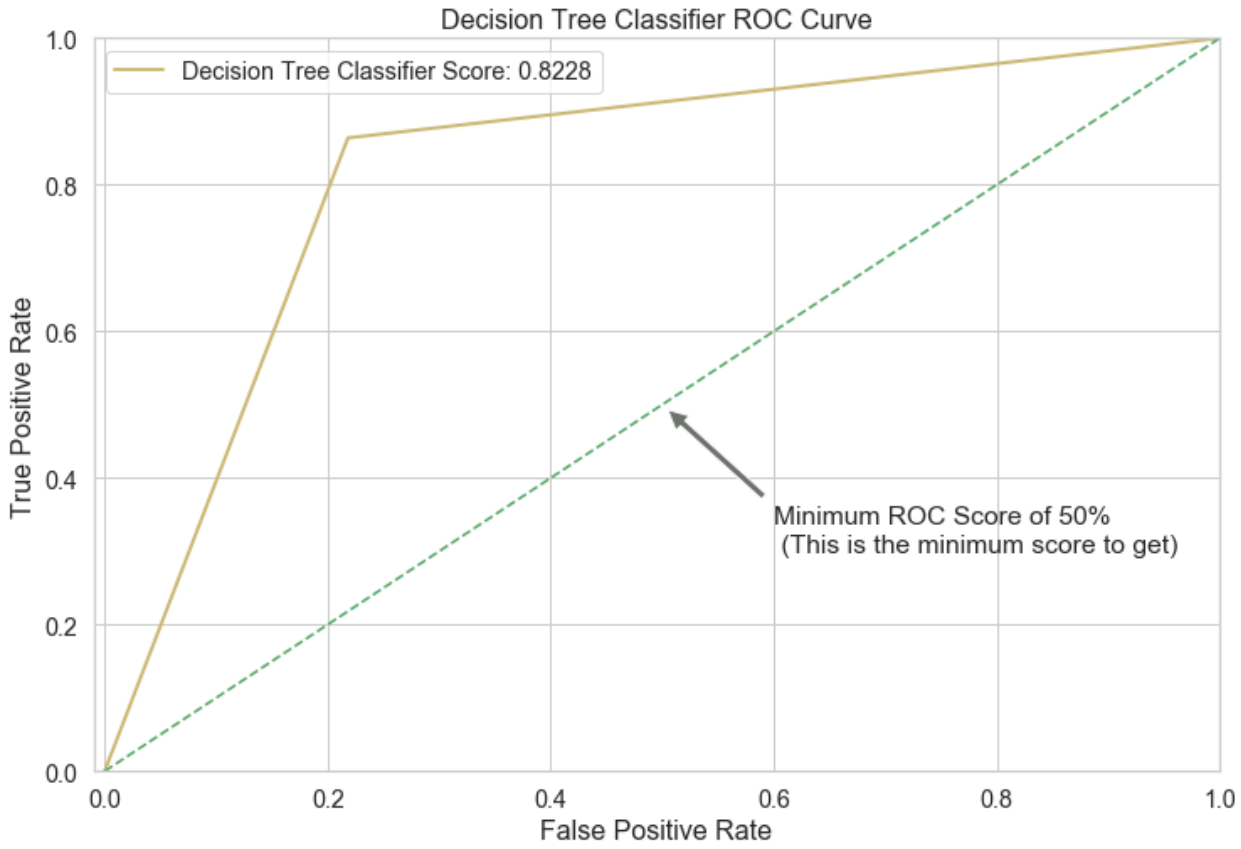


Figure 4.5 Decision Tree Classifier ROC

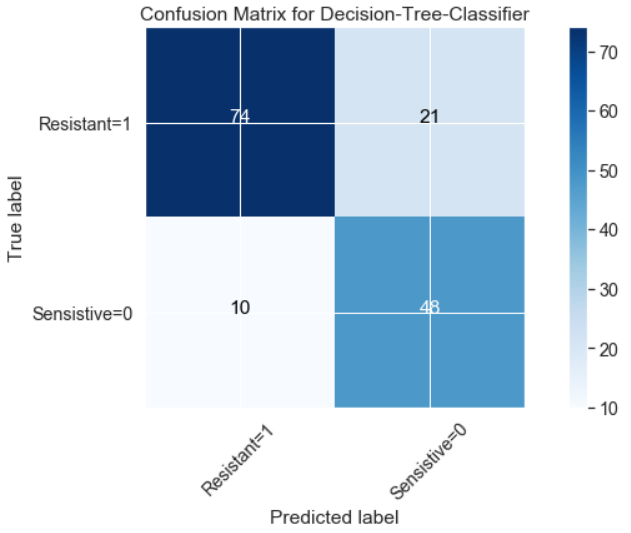


Figure 4.5 Decision Tree Classifier Confusion Matrix

From the confusion matrix Decision Tree Classifier in figure 4.5.4.2, we observe that the True positives are 74, the True negatives are 48, false positives are 10 and the False Negatives are 21.

4.6 Comparison of Machine Learning Models

4.6.1 Classification Reports

Table 4.6 Classification Report

ML Model	Precision Score	Recall Score	F1_Score	Accuracy
Support Vector Machine	0.86	0.9	0.89	0.85
Random Forest	0.85	0.88	0.86	0.9
Logistic Regression	0.83	0.89	0.86	0.88
Decision Tree	0.85	0.86	0.86	0.89

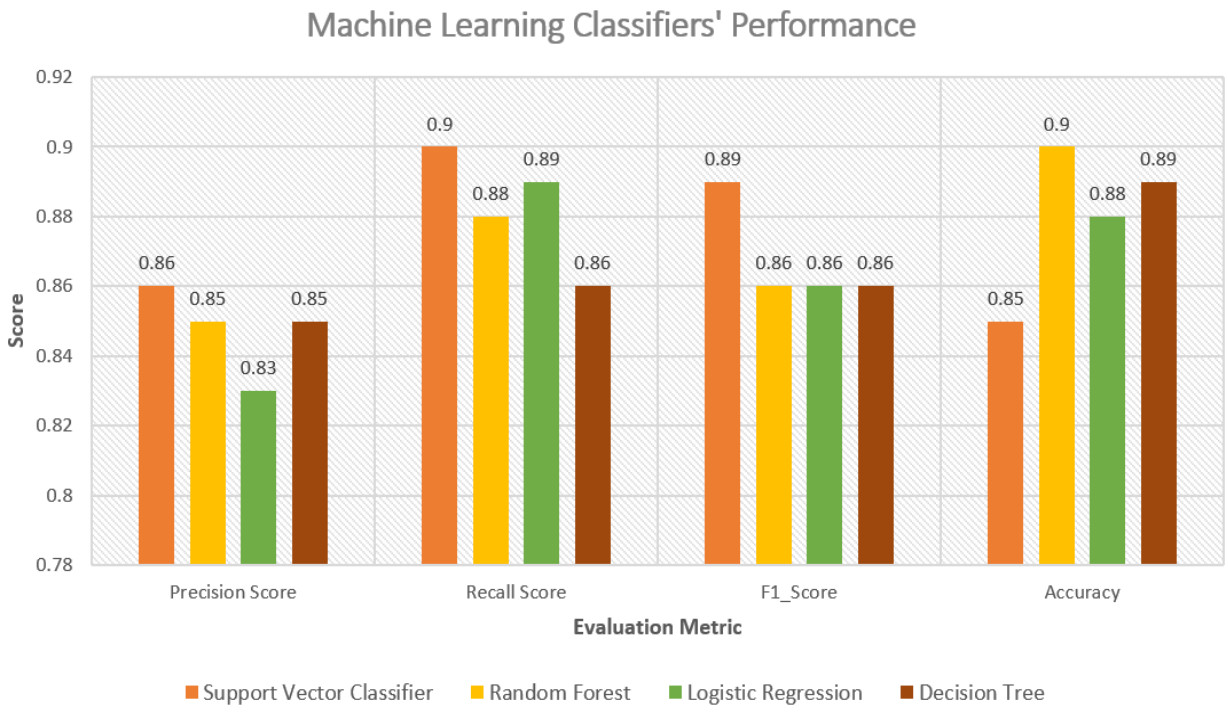


Figure 4.6.1 Machine Learning Classifiers' Performance

It is evident from *table 4.6.1* and figure *4.6.1*, that all the machine-learning models had relatively high evaluation metric scores, however, some models performed better depending on the type of the evaluation metric. Looking at the precision scores, the Support Vector Machine had the highest (0.86). Random Forest and Decision Tree had the same precision of 0.85 while Logistic Regression had the lowest at 0.83. Focusing on the recall, the Support Vector Machine algorithm was found to have the best recall score of 0.9 compared to Logistic Regression (0.89), Random Forest (0.88), and Decision Tree with 0.86. For F1-Score, results reveal that the Support Vector Machine still leads with a score of 0.89, followed by Random Forest, Decision Tree, and Logistic Regression that had the same F1-Score of 0.86. Interestingly, for accuracy, the highest was the Random Forest (0.9) followed by Decision Tree (0.89), Logistic Regression (0.88), and lastly Support Vector Machine (0.85).

4.6.2 ROC Curves

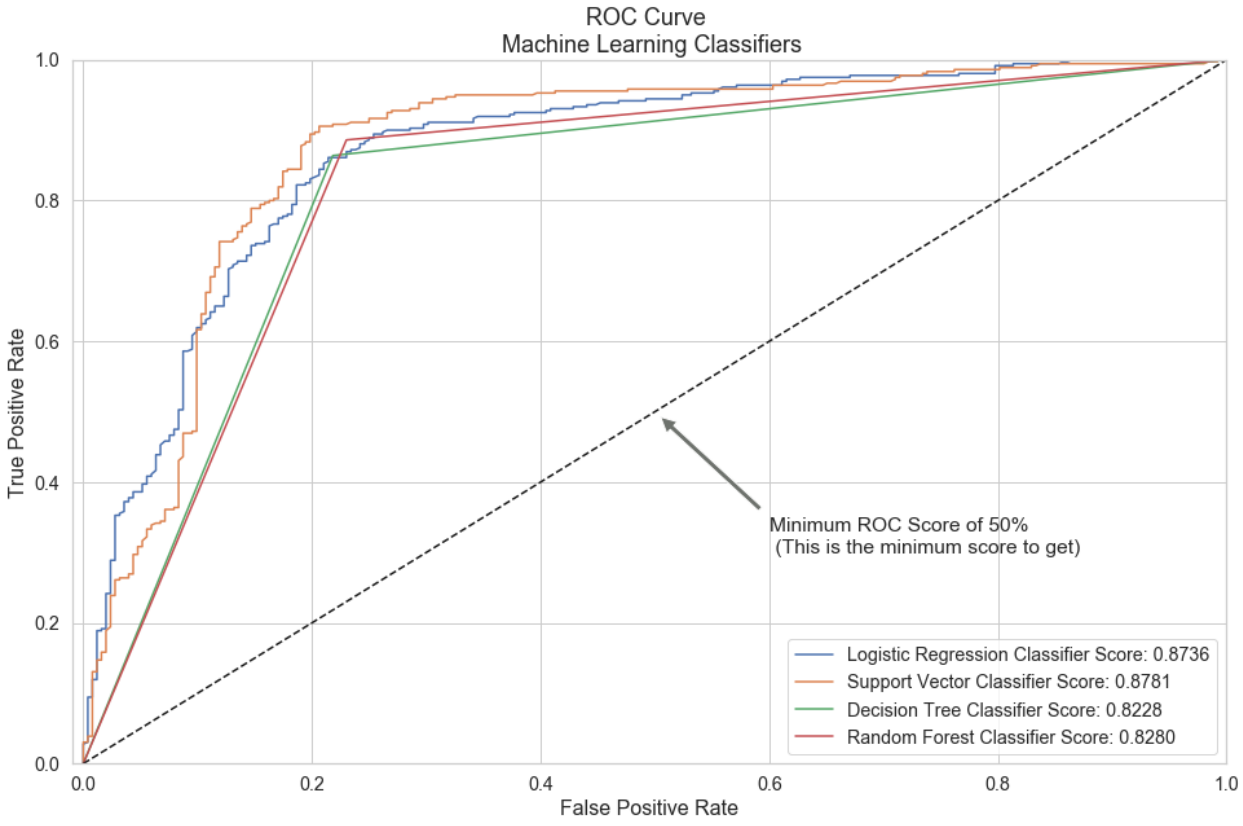


Figure 4.6.2. ROC Curve Machine Learning Classifiers

From ROC Curves in figure 4.6.2 above, it is evident that all the models had an AUC larger than 0.8. Support Vector Machine had the highest AUC (0.8781) followed by Logistic Regression (0.8736). Random Forest and Decision Tree Classifiers had lower but almost the same AUCs of 0.8280 and 0.8228 respectively.

4.6.3 Confusion Matrices

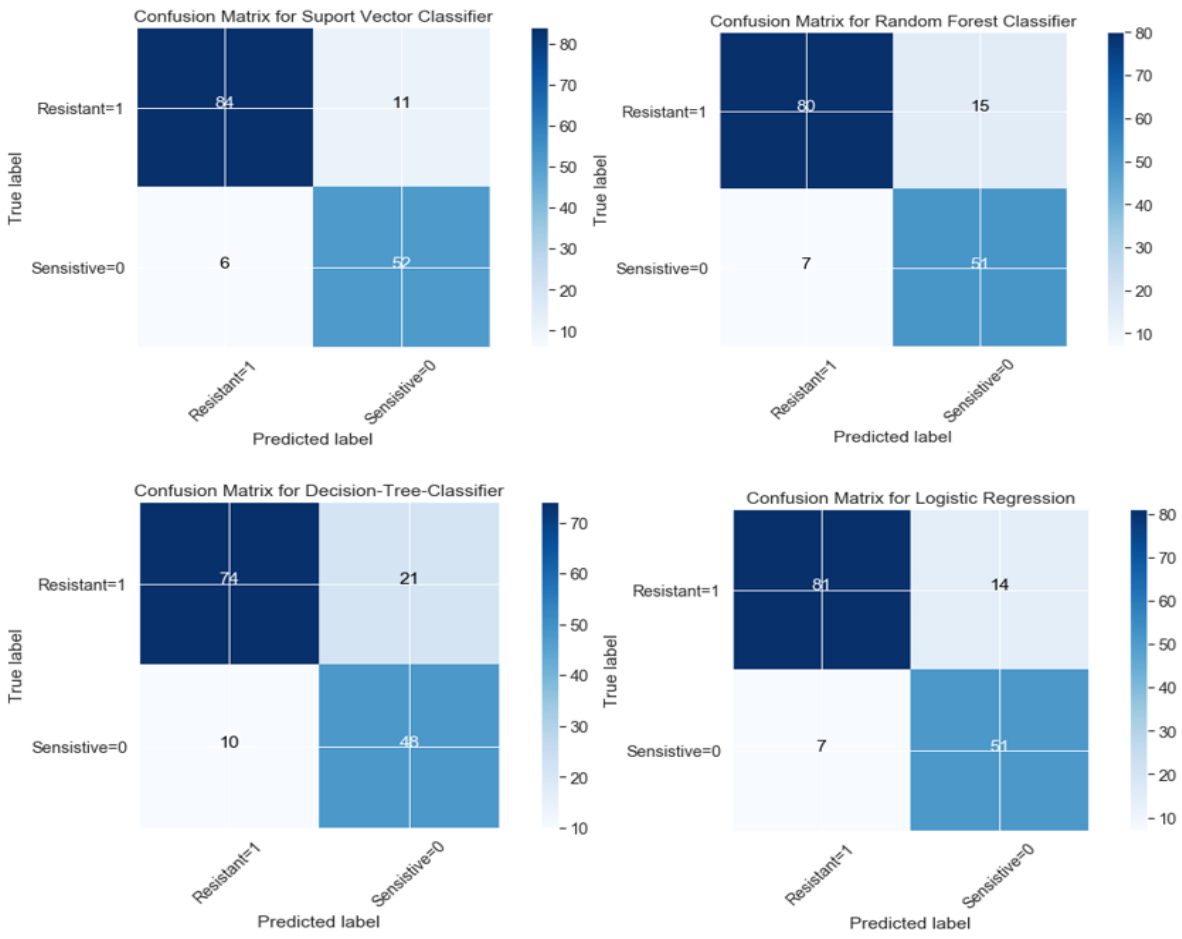


Figure 4.6.3 Confusion Matrices of all the Machine Learning Classifiers

We observe from figure 4.6.3 that Support Vector Machine was able to classify correctly most of the cases (TP=84 and TN=5) Compared to all other models. This also leads the classifier

incorrectly classifying fewer cases than the other models (FP= 6 and FN= 11). Logistic Regression was the second-best classifier when the confusion matrix is considered. It correctly classified fewer cases (TP=81 and TN=51) and incorrectly classified more cases (FP= 7 and FN= 14). This was followed by Random Forest that correctly classified (TP=80 and TN=51) and incorrectly classified (FP= 7 and FN= 15). The least performing was the Decision Tree Classifier that correctly classified (TP=74 and TN=48) and incorrectly classified (FP= 10 and FN= 21)

4.7 Impact of Hyperparameter tuning

The main goal of hyperparameter tuning is to choose a set of optimal **hyperparameters** for a learning algorithm. In this study, it is evident from figures 4.7.1-4.7.4 that there was an improvement in all metric scores of all models after hyperparameter tuning except the Logistic regression's precision that remained constant at 0.83.

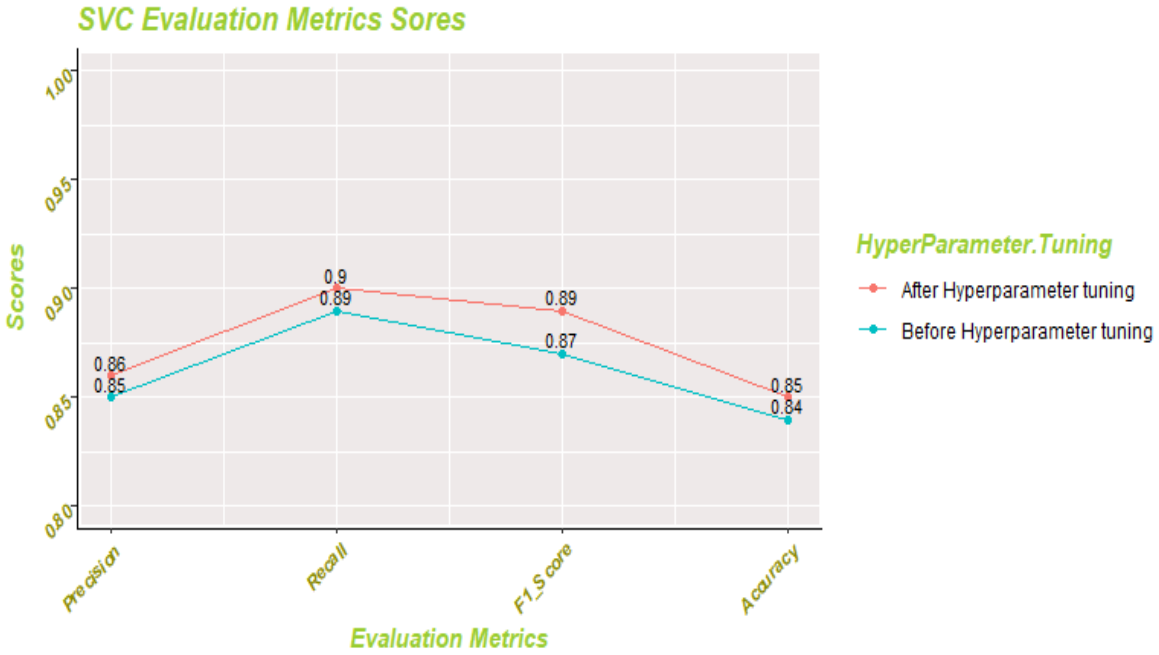


Figure 4.7.1 SVC Evaluation Metrics Scores

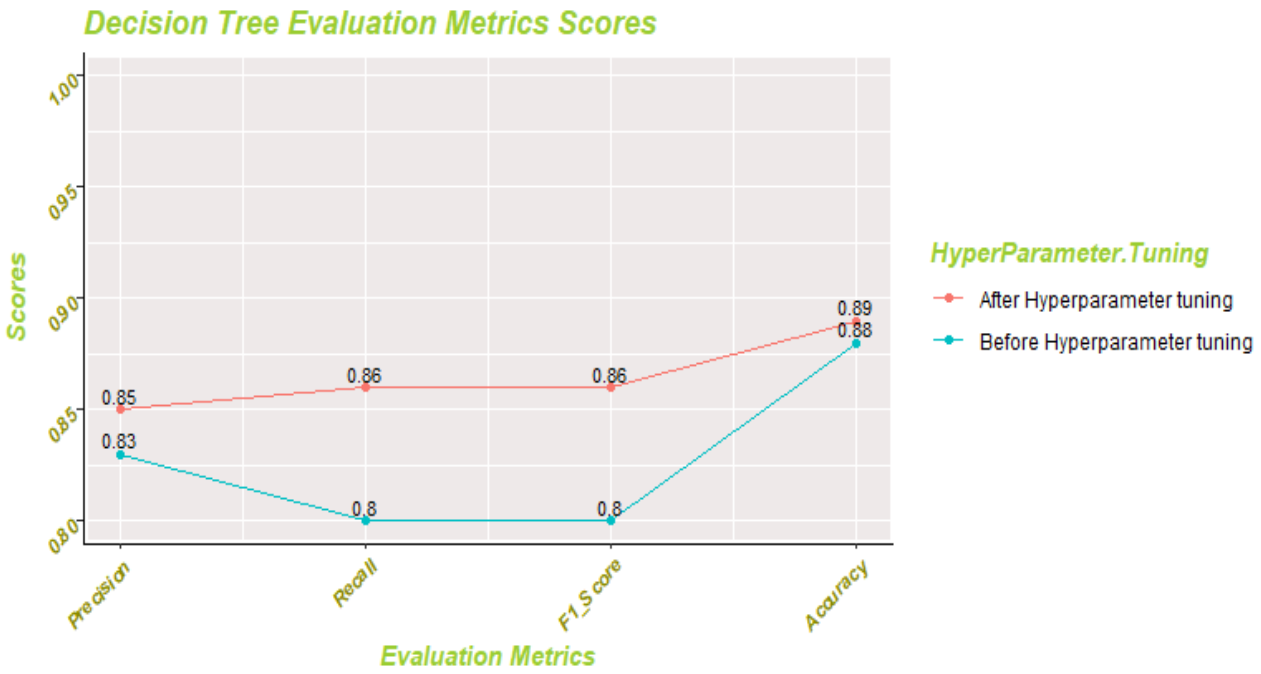


Figure 4.7.2 Decision Tree Evaluation Metrics Scores

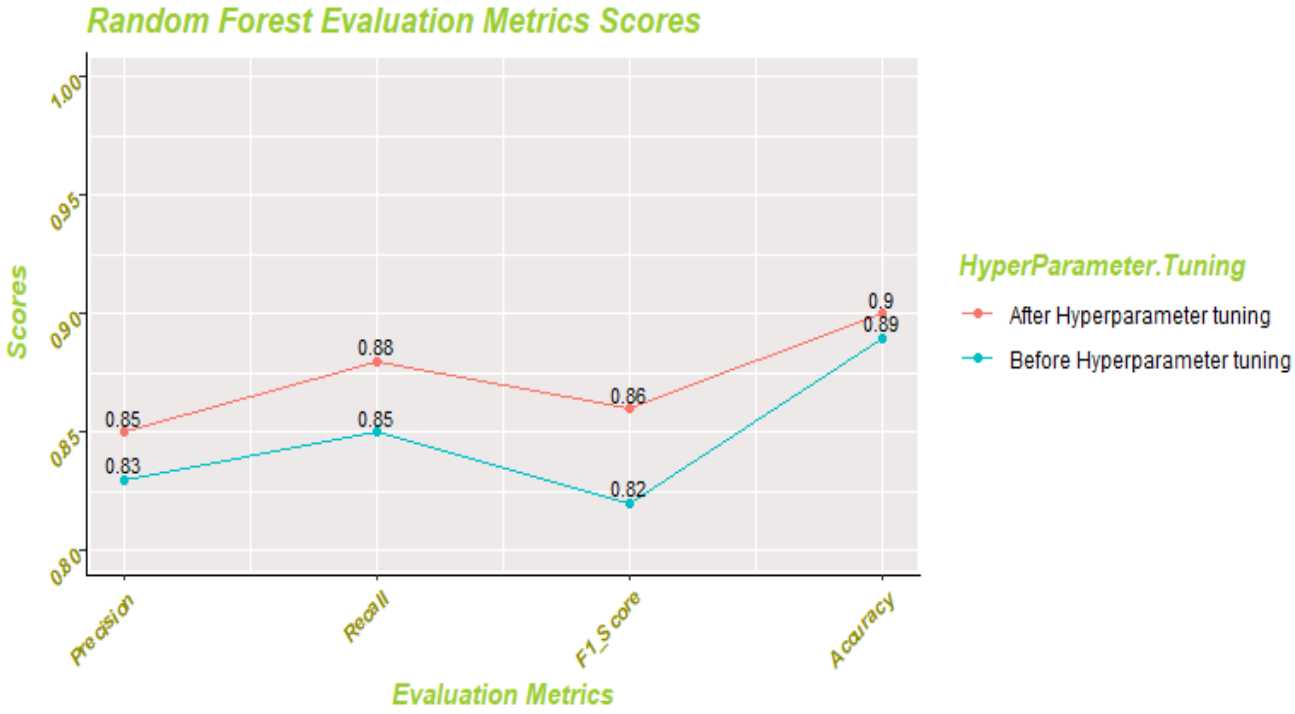


Figure 4.7.3 Random Forest Evaluation Metrics Scores

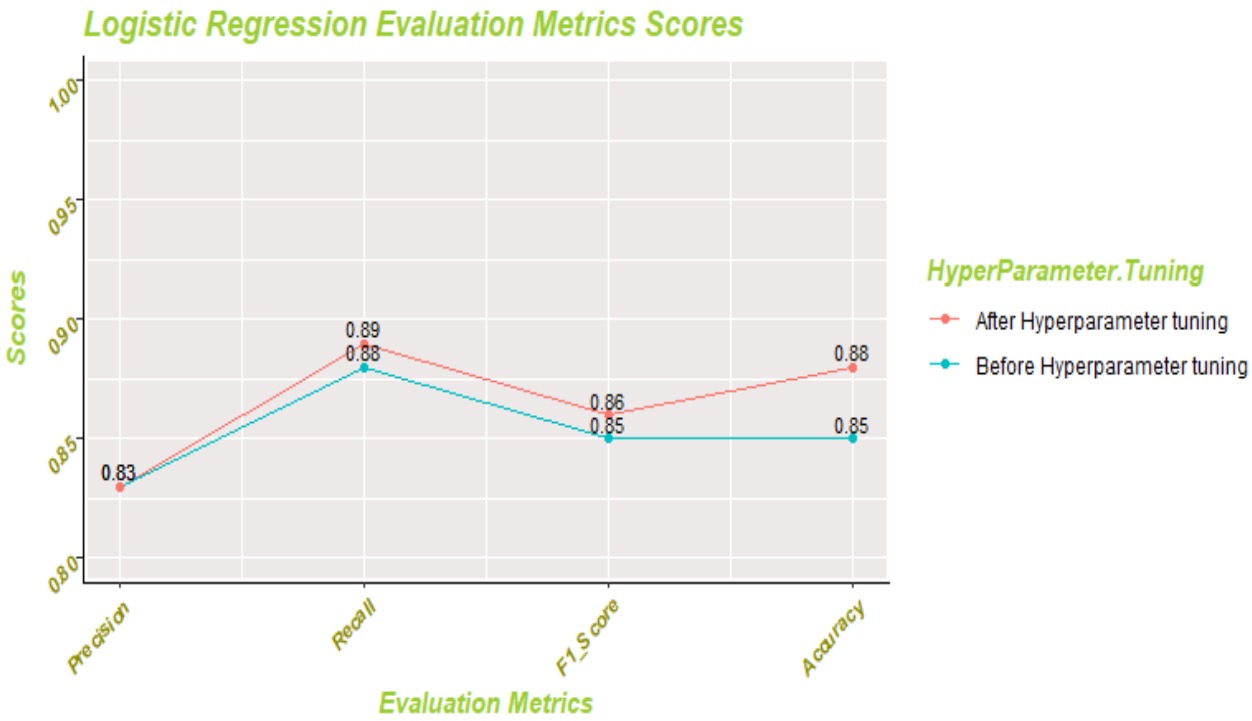


Figure 4.7.4 Logistic Regression Evaluation Metrics Scores

CHAPTER FIVE: DISCUSSION

5.1 Antibiotic resistance patterns and trends to S.Typhi

One of the objectives of this research was to determine the antibiotic resistance pattern for Salmonella Typhi among commonly used antibiotics at a reference laboratory in Rwanda. This study reveals an average resistance to S.Typhi of 131(86.2%) for Cotrimoxazole, 130 (85.5%) for Ampicillin, and 123(80.9%) for chloramphenicol. This reassures that resistance to S.Typhi by the first-line drugs in LMICs is alarming just like [34]. Resistance to ampicillin was generally reducing over the years even though there was an increase in 2017. The trends of chloramphenicol and Cotrimoxazole resistance were sporadic; however, they remained generally high.

Fluoroquinolones have been known to be the most reliable treatment of typhoid fever when the first-line drugs are resistant [8], this study, however, reveals an increasing resistance to ciprofloxacin from 1 (3.6 %) in 2015 to 1 (4.5 %) in 2016, 11(22 %) in 2017, 6 (28.6 %) in 2018 and to 12 (38.7 %) in 2019. Also, Nalidixic acid-resistant (NAL-R) strains are associated with reduced susceptibility to fluoroquinolones. Even though resistance to Nalidixic acid decreased in this study, from 46.4% in 2015 to 18.2% in 2016, a sharp increase to 98% was observed in 2017. Which is indeed a concern. It reduced to 16 (76.2 %) in 2018 and eventually reduced again to 9 (29.0%) in 2019 but then it is still worrying. These findings are in agreement with other studies from LMICs [6]

Third-generation cephalosporins have proved to be an option when fluoroquinolones are resistant. However, this study found that Ceftazidime resistance to S.Typhi was a little high 50 (32.9%) but Cefotaxime seems reliable since its resistance average 11(7.2%), this was low. This is comparable to [19], [43] and [10]. Tetracycline has also proved to be generally reliable since its resistance was low at 15 (9.9%).

5.2 MDR and XDR

Notably, overall multidrug resistance (resistance to chloramphenicol, ampicillin, Cotrimoxazole) over the five years was at 108 (71.1%). This is very high and looking at it annually, 19/ 28 (67.8%) S.Typhi isolates were MDR in 2015 which, decreased in 2016 were 14/22(63.6%), increased drastically to 41/50 (82%) in 2017, then reduced to 14/21 (66.7%) in 2018 and reduced to 20/31

(64.5 %) in 2019. In spite of the decrease, MDR was generally high throughout the five years. It has been found that MDR organisms are widely distributed in both East and West Africa,

Where for instance in Ghana (68/101; 67%), Kenya (50/59; 85%), Tanzania (4/11; 36%), and Uganda (7/30; 23%) [18]. More similar results from studies conducted in LMICs such as [10]and [7] show how MDR is a current problem of interest.

It is indeed more of a concern that XDR (resistance to chloramphenicol, ampicillin, trimethoprim-sulfamethoxazole, fluoroquinolones, and third-generation cephalosporin) cases were found 4/152 (2.6%). Interestingly, they were all in the year 2017. Similarly, reports in 2017 revealed a large outbreak in Sindh, Pakistan of XDR[72] This year may be of interest to look into since getting XDR means limited options of treatment given that the development of new drugs is at a very slow pace. This may be comparable to a study conducted by [40], which calls for the need to carry out genetic diversity of the S.Typhi in different settings and trend in drug resistance and susceptibility

5.3 Age-related patterns of AMR

Salmonella Typhi seems to be most dominant among children as it is evident that the highest number of isolates, 76 (50%) were from children between 0-15 years. Given most 140 (92.1%), of the isolates were from blood, it calls for urgent response on ways to treat Salmonella Typhi patients on time. Similar findings were found by [73] that the primary cause bloodstream infection is *Salmonella* Typhi among children <2 years.

S.Typhi isolates had the highest resistance among those above 30 years. This applies to almost all the antibiotics (Ampicillin (90.5 %), Cefotaxime 2(9.5%), Ceftadizime 8 (38.1%), Ciprofloxacin 31 (20.4 %), Nalidixic acid 91 (59.9 %) and Tetracycline 4 (19 %)) except for chloramphenical (38(84.4 %)) that had the highest resistance among those of 11-20 years. This may imply that being an adult is associated with higher resistance. [74], found similar results after categorizing the S.Typhi isolates in five age categories, <10, 10-20, 20-30, 30-40 and >40 years and used twelve antibiotics. It found out that the rate of resistance was highest among patients of 30-40 years.

S.Typhi resistance to Cefotaxime and Tetracycline was low among all the age groups, 0-10, 10-20, 20-30, and to those above 30 years, thus seem to be reliable antibiotics. Examining antibiotics per age group, Cefotaxime (resistance = 4 (6.7%)) seem the best antibiotic for those between 0-10

years, Tetracycline (resistance = 1(2.2%) to those between 10-20, cefotaxime, ciprofloxacin and tetracycline (resistance = 7(7.7%)) is best to those between 20-30 years while Cefotaxime (resistance=2 (9.5%)) seems to be among those above 30 years.

5.4 Machine Learning model Performance

Another objective of this study involved developing antibiotic resistance predictive models for determining drug susceptibility patterns using Machine Learning Techniques. Based on the metrics used to evaluate the models, we were able to select the best one.

The recall scores (table 4.6.1), tells us the ability of the classifiers to find all the resistant cases. Therefore, we can say that when a patient with S.Typhi is resistant to an antibiotic, the Support Vector Machine can correctly predict that one is resistant with the highest score of 0.9. Logistic Regression correctly predicted that one is resistant given that a patient with S.Typhi is resistant with a score of 0.89. Random Forest can predict that one is resistant by 0.88 from its recall score. The decision tree had the least recall score of 0.86 compared to all other models. This means that it can predict 0.86 times correctly that one is resistant when a patient with S.Typhi is resistant to an antibiotic. We would want to maximize this score and thus we can say Support Vector Machine is the most dominant predictor for any resistance in our model

Focusing on Precision, which in this case, is the ability of the above models not to classify as resistant in a sample that is sensitive; Support Vector Machine still performs best just as it performed with recall score. The score of 0.86 implies that the Support Vector Machine cannot classify a sensitive patient to S.Typhi as resistant 0.86 times. In other words, when a Support Vector Machine predicts that a patient with S.Typhi is resistant to an antibiotic, it is 0.86 times correct. Random forest and Decision tree has the same precision of 0.85, which is a bit lower than that of SVC. This score means that when the two classifiers predict that a patient with S.Typhi is resistant to an antibiotic, they are 0.85 times more likely correct. Logistic regression, on the other hand, tells us that when it predicts that a patient with S.Typhi is resistant to an antibiotic, it is more likely to be 0.83 times correct. In this case, we choose the Support Vector Machine as our best model.

FI_Score on the other hand combines both precision and recall so that we can have a way to interpret both scores using one overall score. Therefore, from our results, the Support Vector Machine stands out to have the highest F1-Score of 0.89, followed by Random Forest, Decision Tree, and Logistic Regression that had the same F1 Score of 0.86. This implies that the Support Vector Machine is the best classification model when F1-Score is used as the evaluation metric.

In this study, looking at classification accuracy, which tells us how best a machine-learning model can identify patterns between variables based on training data, Random Forest seems to have the best accuracy score of 0.9. This implies that Random Forest was able to correctly predict 90% of the input samples that they are resistant to S.Typhi. Decision Tree accuracy score of 0.89, means it could correctly predict a bit lower ratios (89%) of the number of input samples. Logistic Regression accuracy score (0.88) and Support Vector Machine accuracy score (0.85) means Logistic Regression could correctly predict 88% of all the input samples that they are resistant while Support Vector Machine correctly predicted 85% of the input samples that they are resistant to S.Typhi.

Usually, the further the curve is to the top left corner, the better the model's performance. However, sometimes it is difficult to know exactly which one has the largest area and thus the AUC helps us to know the exact area under the ROC Curve of each machine-learning model and thus performance can be evaluated. From the results (figure 4.6.2), it is evident that the Support Vector Machine has the highest AUC (0.8781), this implies that it is the best performing classifier followed by Logistic Regression (AUC= 0.8741). Random Forest had a lower performance AUCs of 0.8248 while Decision Tree Classifiers had the lowest performance due to its AUC of 0.8248. A similar study conducted by [55] gave high AUC.

The Confusion Matrices also contributed a lot in determining the quality of the output of the entire Machine learning Classifiers. From figure 4.6.3, it is clear that the Support Vector Machine had the best performance since it was able to classify correctly most of the cases and incorrectly classifying the fewest cases than the other models. The True Positives of 84 for Support Vector Machine, 81 for Logistic Regression, 80 for Random Forest, and 74 for Decision Tree implies that the classifiers could correctly predict 84, 81, 80, and 74 cases respectively that a patient with S.Typhi is resistant to an antibiotic. The True Negative values on the other hand, for Support Vector Machine (52), Logistic Regression (51), Random Forest (51), Decision Tree (48) implies

that Support Vector Machine correctly predicted most cases that the patient with S.Typhi is sensitive to a given antibiotic followed by Logistic Regression then Random Forest and lastly Decision Tree. Focusing on the incorrectly predicted cases and starting with the False Positives, Support Vector Machine had the least, 6 cases followed by Logistic Regression (7) and Random Forest (7) while Decision Tree (10) had the most cases of False Positive. It means that the classifiers incorrectly predicted 6, 7, 7, and 10 cases for Support Vector Machine, Logistic Regression, Random Forest, and Decision Tree respectively that a patient is resistant to a given antibiotic when in fact is sensitive. Secondly, False positive values for Support Vector Machine, Logistic Regression, Random Forest and Decision Tree (11, 14, 15, and 21 respectively) implied that Support Vector Machine incorrectly predicted the least number of cases (11) that a patient is sensitive to an antibiotic when in fact is resistant.

Therefore, based on these performance results, the Support Vector Machine was selected as the best model for this study. Even though its accuracy of 0.85 was slightly lower, compared to other classifiers, it achieved an outstanding performance based on the confusion matrix since it was able to classify correctly most of the cases and incorrectly classifying the fewest cases than the other models. Moreover, it had the highest recall of 0.9, a precision of (0.86), F1_Score of 0.89, and Area Under the ROC Curve of 0.8781. Support Vector Machine's low accuracy of 0.85 means that the model incorrectly classified 15% of the instances to some other classes. The possible reason may be due to the nature of the dataset used in this study.

This study did not include clinical information. Even though using clinical information could improve the performance of the models, it may be expensive especially when the results are needed on time. Performing models were able to be built from this research by only using the existence of data from the Microbiology department in the Laboratory without the clinical information.

CHAPTER SIX: CONCLUSION AND RECOMMENDATIONS

6.1 Conclusion

The findings from this study suggest that machine learning techniques can be a powerful tool in combating global antimicrobial drug resistance threat on time through high prediction accuracy in predicting antibiotic resistance.

In this study, we demonstrate that antibiotic resistance models trained alongside hyperparameter tuning, we can achieve high F1-Score and accuracy in the detection of antibiotic resistance. Integrating Artificial Intelligence and Machine learning techniques to large Antimicrobial Resistance datasets can be used to build predictive AMR models that can be used to serve patients even in the ICU. Ultimately, such predictive models can be utilized to enhance both national surveillance of antimicrobial drug resistance and for planning purposes.

6.2 Study Strengths and Limitations

This study stood out because it was able to apply data science by introducing the use of artificial interlligence (AI) through machine learning algorithms to predicting drug resitance in S.Typhi using the available data without any clinical information. This can save lives in the intensive care unit when time is limited[16].

This study was, however limited by data sample size in that, the number of isolates was small and even the features for predicting AMR. However, k-fold cross validation helped to improve the performance of the model. Even though including variables like patient conditions, previous culture history, and other clinical data in a machine learning model to predict susceptibility prove to be efficient [28], such data may sometimes be limited and using the available data is still sufficient to get high prediction accuracy[16]

6.3 Future Work and Recommendation

The data set used in this study was of five years (2015-2019). It was used in checking patterns through descriptive statistics and also predicting antibiotic resistance. In the future, a larger dataset of up to ten years can be used in prediction and by applying time series machine learning algorithms to compare patterns through time.

In this study, factors like genes, drug usage behavior, and clinical information like inpatient/outpatient were not considered. This is because the retrospective data set that was

available did not have such information and due to financial and time constraints a prospective could not be conducted. Usually, these are very important variables that can affect antibiotic resistance. Therefore, future heath data science can include these variables to predict antibiotic resistance.

This research focused on only four algorithms: Support Vector Machine, Decision Trees, Random Forest, and Logistic regression however there are other machine learning classifiers such as XGboost and GradientBoosting that may be used to produce a good performance in predicting antibiotic resistance. Also, the k-fold cross-validation method was used as an evaluation method, but there exist other resampling methods like leave one out cross-validation. The evaluation metrics used in this study were accuracy, precision, recall, confusion matrix, and ROC Curve. Other metrics like log-loss can be used in the future.

Finally, different types of bacteria can be used to predict resistance, unlike this study that considered only laboratory-based datasets for S.Typhi.

REFERENCES

[1] F. N. Qamar *et al.*, "Outbreak investigation of ceftriaxone-resistant Salmonella enterica serotype Typhi and its risk factors among the general population in Hyderabad, Pakistan: a matched case-control study," *Lancet Infect. Dis.*, vol. 18, no. 12, pp. 1368–1376, 2018.

[2] A. H. Havelaar *et al.*, "World Health Organization Global Estimates and Regional Comparisons of the Burden of Foodborne Disease in 2010," *PLoS Med.*, vol. 12, no. 12, pp. 1–23, 2015.

[3] M. Antillón *et al.*, "The burden of typhoid fever in low- and middle-income countries: A meta-regression approach," *PLoS Negl. Trop. Dis.*, vol. 11, no. 2, pp. 1–21, 2017.

[4] T. Hampton, "Report reveals scope of US antibiotic resistance threat," *JAMA - J. Am. Med. Assoc.*, vol. 310, no. 16, pp. 1661–1663, 2013.

[5] P. Dadgostar, "Antimicrobial resistance: implications and costs," *Infect. Drug Resist.*, vol. 12, pp. 3903–3910, 2019.

[6] A. J. Browne *et al.*, "Drug-resistant enteric fever worldwide, 1990 to 2018: A systematic review and meta-analysis," *BMC Med.*, vol. 18, no. 1, pp. 1–22, 2020.

[7] C. D. Britto, V. K. Wong, G. Dougan, and A. J. Pollard, "A systematic review of antimicrobial resistance in Salmonella enterica serovar Typhi, the etiological agent of typhoid," *PLoS Negl. Trop. Dis.*, vol. 12, no. 10, pp. 1–15, 2018.

[8] B. T. Tadesse *et al.*, "Antimicrobial resistance in Africa: A systematic review," *BMC Infect. Dis.*, vol. 17, no. 1, pp. 1–17, 2017.

[9] K. O. Akinyemi *et al.*, "Typhoid fever: Tracking the trend in Nigeria," *Am. J. Trop. Med. Hyg.*, vol. 99, no. 3, pp. 41–47, 2018.

[10] W. C. Mutai, A. W. T. Muigai, P. Waiyaki, and S. Kariuki, "Multi-drug resistant Salmonella enterica serovar Typhi isolates with reduced susceptibility to ciprofloxacin in Kenya," *BMC Microbiol.*, vol. 18, no. 1, pp. 4–8, 2018.

[11] T. Mashe *et al.*, "Laboratory characterisation of Salmonella enterica serotype Typhi isolates

from Zimbabwe, 2009-2017," *BMC Infect. Dis.*, vol. 19, no. 1, pp. 1–9, 2019.

[12] S. E. M. Journal, "Archive of SID Antimicrobial Susceptibility Patterns of Salmonella Typhi From Ki- Archive of SID," vol. 11, no. 3, pp. 117–121, 2010.

[13] Clinical and Laboratory Standards Institute, *Performance standards for antimicrobial disk susceptibility tests: Approved standard - Eleventh edition*, vol. 32, no. 1. 2012.

[14] J. Hudzicki, "Kirby-Bauer Disk Diffusion Susceptibility Test Protocol Author Information," *Am. Soc. Microbiol.*, no. December 2009, pp. 1–13, 2016.

[15] M. W. Pesesky *et al.*, "Evaluation of machine learning and rules-based approaches for predicting antimicrobial resistance profiles in gram-negative bacilli from whole genome sequence data," *Front. Microbiol.*, vol. 7, no. NOV, pp. 1–17, 2016.

[16] G. Feretzakis *et al.*, "Using machine learning techniques to aid empirical antibiotic therapy decisions in the intensive care unit of a general hospital in Greece," *Antibiotics*, vol. 9, no. 2, 2020.

[17] H. R. Karim, M. Yunus, and P. Bhattacharyya, "A retrospective study of endotracheal or tracheostomy tube blockage and their impact on the patients in an intensive care unit," *J. Mahatma Gandhi Inst. Med. Sci.*, vol. 22, no. 1, p. 12, 2017.

[18] S. E. Park *et al.*, "The phylogeography and incidence of multi-drug resistant typhoid fever in sub-Saharan Africa," *Nat. Commun.*, vol. 9, no. 1, 2018.

[19] B. Hasan, S. G. Nahar, L. Akter, and A. A. Saleh, "Antimicrobial sensitivity pattern of Salmonella typhi isolated from blood culture in a referral hospital," *Bangladesh J. Med. Microbiol.*, vol. 5, no. 1, pp. 16–20, 2011.

[20] C. D. Britto, V. K. Wong, G. Dougan, and A. J. Pollard, "A systematic review of antimicrobial resistance in Salmonella enterica serovar Typhi, the etiological agent of typhoid," *PLoS Negl. Trop. Dis.*, vol. 12, no. 10, pp. 1–15, 2018.

[21] H. Schatten and A. Eisenstark, "Salmonella: Methods and protocols: Second edition," *Salmonella Methods Protoc. Second Ed.*, vol. 1225, pp. 1–287, 2014.

[22] J. Gatabazi, "P085: Surveillance of drug-resistant Salmonella sp and Shigella sp infections

in Rwanda," *Antimicrob. Resist. Infect. Control*, vol. 2, no. S1, p. 2013, 2013.

[23] K. Syal *et al.*, "Current and emerging techniques for antibiotic susceptibility tests," *Theranostics*, vol. 7, no. 7, pp. 1795–1805, 2017.

[24] M. Carroll, A. Rangaiahagari, E. Musabeyezu, D. Singer, and O. Ogbuagu, "Five-year antimicrobial susceptibility trends among bacterial isolates from a tertiary health-care facility in Kigali, Rwanda," *Am. J. Trop. Med. Hyg.*, vol. 95, no. 6, pp. 1277–1283, 2016.

[25] J. M. Monk, "Predicting antimicrobial resistance and associated genomic features from whole-genome sequencing," *J. Clin. Microbiol.*, vol. 57, no. 2, pp. 1–4, 2019.

[26] S. Martínez-Agüero, I. Mora-Jiménez, J. Lérida-García, J. Álvarez-Rodríguez, and C. Soguero-Ruiz, "Machine learning techniques to identify antimicrobial resistance in the intensive care unit," *Entropy*, vol. 21, no. 6, pp. 1–24, 2019.

[27] M. Oonsivilai *et al.*, "Using machine learning to guide targeted and locally-tailored empiric antibiotic prescribing in a children's hospital in Cambodia [version 1; referees: 2 approved]," *Wellcome Open Res.*, vol. 3, no. 0, pp. 1–18, 2018.

[28] D. Ghosh *et al.*, "Machine learning based prediction of antibiotic sensitivity in patients with critical illness," *medRxiv*, p. 19007153, 2019.

[29] "Could Artificial Intelligence Fix Antibiotic Resistance? - The Atlantic." [Online]. Available: https://www.theatlantic.com/science/archive/2020/03/could-artificial-intelligence-fix-antibiotic-resistance/607978/. [Accessed: 14-Sep-2020].

[30] J. Lv, S. Deng, and L. Zhang, "A review of artificial intelligence applications for antimicrobial resistance," *Biosaf. Heal.*, 2020.

[31] M. Abbas *et al.*, "Conflicts of interest in infection prevention and control research : no smoke without fire . A narrative review," *Intensive Care Med.*, vol. 44, no. 10, pp. 1679–1690, 2018.

[32] S. Omulo, S. M. Thumbi, M. K. Njenga, and D. R. Call, "A review of 40 years of enteric antimicrobial resistance research in Eastern Africa: What can be done better?," *Antimicrob. Resist. Infect. Control*, vol. 4, no. 1, pp. 1–13, 2015.

[33] A. Makkar *et al.*, "Epidemiological Profile and Antimicrobial Resistance Pattern of Enteric Fever in a Tertiary Care Hospital of North India - a Seven Year Ambispective Study," *Acta medica (Hradec Kral.*, vol. 61, no. 4, pp. 125–130, 2018.

[34] A. C. Stark, "Treatment of typhoid fever," *Br. Med. J.*, vol. 2, no. 2347, p. 1677, 1905.

[35] K. A. Date *et al.*, "Changing Patterns in Enteric Fever Incidence and Increasing Antibiotic Resistance of Enteric Fever Isolates in the United States, 2008–2012," vol. 63, no. 3, pp. 322–329, 2017.

[36] A. Ali, H. A. Ali, F. H. Shah, A. Zahid, H. Aslam, and B. Javed, "Pattern of antimicrobial drug resistance of Salmonella Typhi and Paratyphi A in a Teaching Hospital in Islamabad," vol. 67, no. 3, pp. 3–5, 2017.

[37] W. C. Mutai, A. W. T. Muigai, P. Waiyaki, and S. Kariuki, "Multi-drug resistant Salmonella enterica serovar Typhi isolates with reduced susceptibility to ciprofloxacin in Kenya," pp. 4–8, 2018.

[38] K. A. Date *et al.*, "Changing Patterns in Enteric Fever Incidence and Increasing Antibiotic Resistance of Enteric Fever Isolates in the United States, 2008-2012," *Clin. Infect. Dis.*, vol. 63, no. 3, pp. 322–329, 2016.

[39] R. J. Leggiadro, "Emergence of extensively drug-resistant salmonella typhi infections among travelers to or from Pakistan-United States, 2016-2018," *Pediatr. Infect. Dis. J.*, vol. 38, no. 6, p. 630, 2019.

[40] E. J. Klemm *et al.*, "Emergence of an extensively drug-resistant Salmonella enterica serovar typhi clone harboring a promiscuous plasmid encoding resistance to fluoroquinolones and third-generation cephalosporins," *MBio*, vol. 9, no. 1, pp. 1–10, 2018.

[41] A. García-Fernández *et al.*, "Emergence of ciprofloxacin-resistant Salmonella enterica serovar Typhi in Italy," *PLoS One*, vol. 10, no. 6, pp. 1–8, 2015.

[42] F. N. Qamar *et al.*, "A Retrospective Study of Laboratory-Based Enteric Fever Surveillance, Pakistan, 2012-2014," *J. Infect. Dis.*, vol. 218, no. Suppl 4, pp. S201–S205, 2018.

[43] H. Khatun *et al.*, "Clinical profile, antibiotic susceptibility pattern of bacterial isolates and

factors associated with complications in culture-proven typhoid patients admitted to an urban hospital in Bangladesh," *Trop. Med. Int. Heal.*, vol. 23, no. 4, pp. 359–366, 2018.

[44] B. Poonia *et al.*, "Typhoidal Salmonella and Emerging Resistance in Outbreak Proportions," *Int. J. Travel Med. Glob. Heal.*, vol. 6, no. 2, pp. 64–68, 2018.

[45] A. Hussain *et al.*, "Typhoidal Salmonella strains in Pakistan: an impending threat of extensively drug-resistant Salmonella Typhi," *Eur. J. Clin. Microbiol. Infect. Dis.*, vol. 38, no. 11, pp. 2145–2149, 2019.

[46] G. S. Laghari, Z. Hussain, S. Z. M. Hussain, H. Kumar, S. M. M. Uddin, and A. Haq, "Antimicrobial Susceptibility Patterns of Salmonella Species in Southern Pakistan," *Cureus*, vol. 11, no. 4, 2019.

[47] J. R. Quinlan, "Induction of decision trees," *Mach. Learn.*, vol. 1, no. 1, pp. 81–106, 1986.

[48] S. M. Kavai and S. Kariuki, "Increasing multidrug and fluoroquinolone resistance among Salmonella Typhi from sporadic outbreaks in Kenya," *Int. J. Infect. Dis.*, vol. 79, no. 2019, p. 44, 2019.

[49] A. Makkar *et al.*, "Epidemiological Profile and Antimicrobial Resistance Pattern of Enteric Fever in a Tertiary Care Hospital of North India - a Seven Year Ambispective Study," *Acta medica (Hradec Kral.*, vol. 61, no. 4, pp. 125–130, 2018.

[50] M. BisiJohnson *et al.*, "Can herbal remedies be the answer to multidrug resistance? Profile of drug resistance in <i>Salmonella</i> species in Eastern Cape, South Africa," *J. Exp. Integr. Med.*, vol. 2, no. 2, p. 147, 2012.

[51] "Elements of Antimicrobial Stewardship." [Online]. Available: https://www.treatsystems.com/Elements. [Accessed: 07-Sep-2020].

[52] D. A. Clifton, K. E. Niehaus, P. Charlton, and G. W. Colopy, "Health Informatics via Machine Learning for the Clinical Management of Patients," *Yearb. Med. Inform.*, vol. 10, no. 1, pp. 38–43, 2015.

[53] M. Su, S. W. Satola, and T. D. Read, "Genome-based prediction of bacterial antibiotic resistance," *J. Clin. Microbiol.*, vol. 57, no. 3, pp. 1–15, 2019.

[54] K. E. Niehaus, T. M. Walker, D. W. Crook, T. E. A. Peto, and D. A. Clifton, “Machine learning for the prediction of antibacterial susceptibility in Mycobacterium tuberculosis,” *2014 IEEE-EMBS Int. Conf. Biomed. Heal. Informatics, BHI 2014*, pp. 618–621, 2014.

[55] M. L. Chen *et al.*, “Beyond multidrug resistance: Leveraging rare variants with machine and statistical learning models in Mycobacterium tuberculosis resistance prediction,” *EBioMedicine*, vol. 43, pp. 356–369, 2019.

[56] D. Moradigaravand, M. Palm, A. Farewell, V. Mustonen, J. Warringer, and L. Parts, “Prediction of antibiotic resistance in Escherichia coli from large-scale pan-genome data,” *PLoS Comput. Biol.*, vol. 14, no. 12, pp. 1–17, 2018.

[57] Z. Liu *et al.*, “Evaluation of Machine Learning Models for Predicting Antimicrobial Resistance of Actinobacillus pleuropneumoniae From Whole Genome Sequences,” *Front. Microbiol.*, vol. 11, no. February, pp. 1–7, 2020.

[58] S. W. Long, H. M. Hospital, P. F. Mcdermott, and U. S. Food, “crossm Using Machine Learning To Predict Antimicrobial MICs and,” no. October, 2019.

[59] I. Yelin *et al.*, “tract infections,” vol. 25, no. 7, pp. 1143–1152, 2020.

[60] M. I. Jordan and T. M. Mitchell, “Machine learning: Trends, perspectives, and prospects,” *Science (80-.).*, vol. 349, no. 6245, pp. 255–260, 2015.

[61] A. Garcia, T. Delorme, and P. Nasr, “Patient age as a factor of antibiotic resistance in methicillinresistant staphylococcus aureus,” *J. Med. Microbiol.*, vol. 66, no. 12, pp. 1782–1789, 2017.

[62] A. Murugappan, J. S. Sudarsan, and A. Manoharan, “Effects of using Lignite mine drainage for irrigation on soils - A case study of perumal tank command area in Tamilnadu State,” *J. Ind. Pollut. Control*, vol. 22, no. 1, pp. 149–160, 2006.

[63] L. R. Kolozsvári *et al.*, “Patient-related factors, antibiotic prescribing and antimicrobial resistance of the commensal Staphylococcus aureus and Streptococcus pneumoniae in a healthy population - Hungarian results of the APRES study,” *BMC Infect. Dis.*, vol. 19, no. 1, pp. 1–8, 2019.

[64] R. Konieczny and R. Idczak, “Mössbauer study of Fe-Re alloys prepared by mechanical alloying,” *Hyperfine Interact.*, vol. 237, no. 1, pp. 1–8, 2016.

[65] “Decision Tree Split Methods | Decision Tree Machine Learning.” [Online]. Available: https://www.analyticsvidhya.com/blog/2020/06/4-ways-split-decision-tree/. [Accessed: 07-Sep-2020].

[66] A. L. and M. Wiener, “Classification and Regression by randomForest. R News 2,” vol. 3, no. December 2002, pp. 18–22, 2003.

[67] L. Breiman, “No Title,” pp. 1–33, 2001.

[68] “Decision Tree vs. Random Forest - Which Algorithm Should you Use?” [Online]. Available: https://www.analyticsvidhya.com/blog/2020/05/decision-tree-vs-random-forest-algorithm/. [Accessed: 07-Sep-2020].

[69] “Practical Guide to Logistic Regression Analysis in R Tutorials & Notes | Machine Learning | HackerEarth.” [Online]. Available: https://www.hackerearth.com/practice/machine-learning/machine-learning-algorithms/logistic-regression-analysis-r/tutorial/. [Accessed: 07-Sep-2020].

[70] “Support Vector Machine Tutorial — (SVM) | Data Science and Machine Learning | Kaggle.” [Online]. Available: https://www.kaggle.com/getting-started/130883. [Accessed: 07-Sep-2020].

[71] “3.1. Cross-validation: evaluating estimator performance — scikit-learn 0.23.2 documentation.” [Online]. Available: https://scikit-learn.org/stable/modules/cross_validation.html. [Accessed: 07-Sep-2020].

[72] Z. A. Dyson, E. J. Klemm, S. Palmer, and G. Dougan, “Antibiotic resistance and typhoid,” *Clin. Infect. Dis.*, vol. 68, no. Suppl 2, pp. S165–S170, 2019.

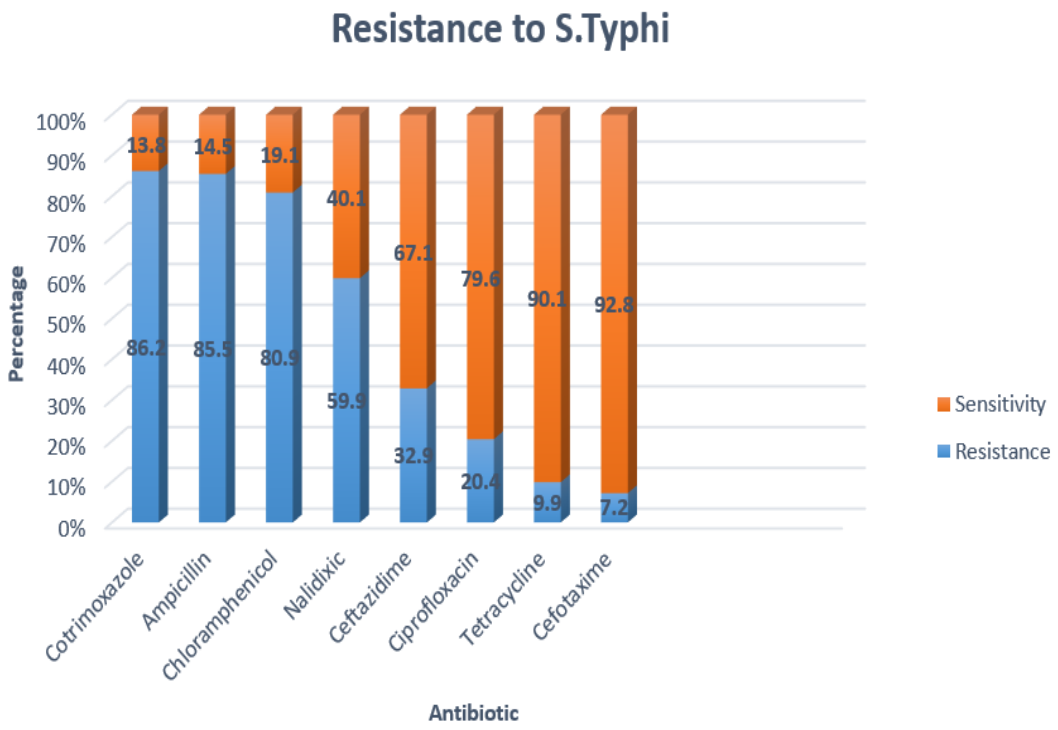
[73] O. A. Msemo *et al.*, “Epidemiology and antimicrobial susceptibility of salmonella enterica bloodstream isolates among febrile children in a rural district in Northeastern Tanzania: A cross-sectional study,” *Clin. Infect. Dis.*, vol. 68, no. Suppl 2, pp. S177–S182, 2019.

[74] A. Mannan, M. Shohel, S. Rajia, N. U. Mahmud, S. Kabir, and I. Hasan, “A cross sectional

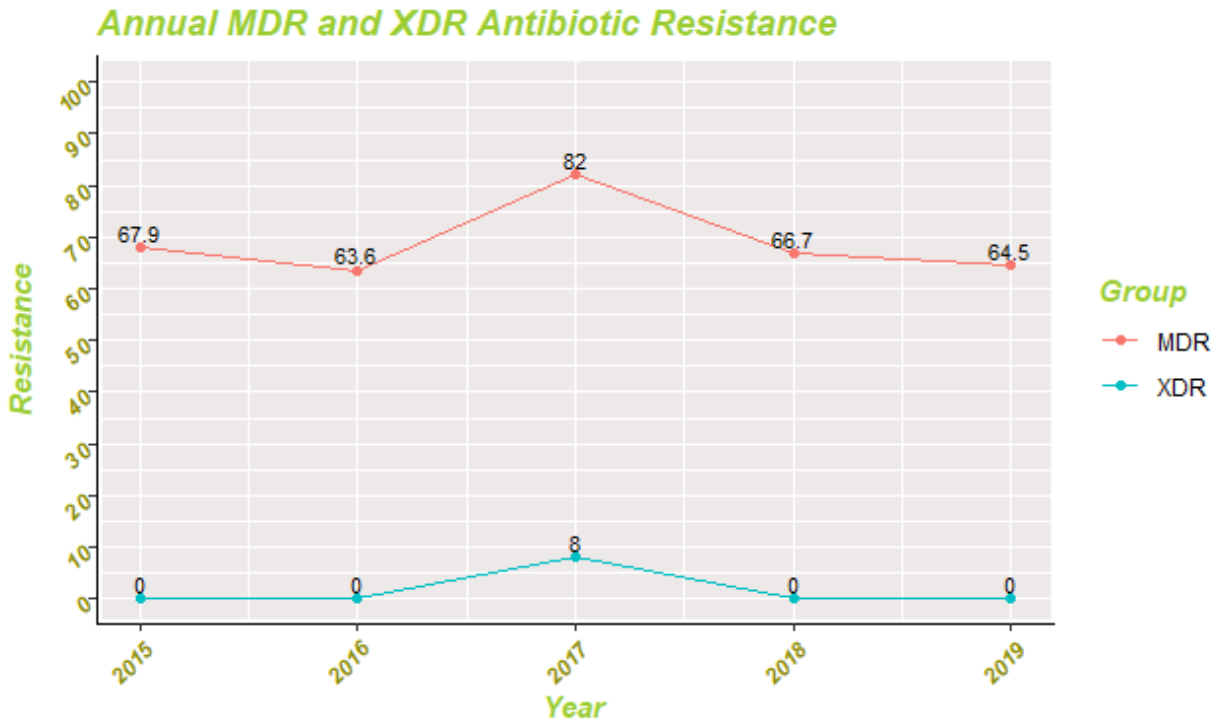
study on antibiotic resistance pattern of Salmonella typhi clinical isolates from Bangladesh," *Asian Pac. J. Trop. Biomed.*, vol. 4, no. 4, pp. 306–311, 2014.

APPENDICE 1

Overall patterns of Antibiotic Resistance to S.Typhi



APPENDICE 2


Annual MDR and XDR Antibiotic Resistance
Resistance
Year
Group
MDR
XDR
67.9
63.6
82
66.7
64.5
0
0
8
0
0
2015
2016
2017
2018
2019